Psychological Test and Assessment Modeling, Volume 61, 2019 (1), 13-36

The impacts of characteristics of disconnected subsets on group anchoring in incomplete rater-mediated assessment networks

Stefanie A. Wind[1] *& Catanya G. Stager*[2]

Abstract

In operational administrations of rater-mediated performance assessments, practical constraints often result in incomplete data collection designs, in which each rater does not rate each performance on each task. Unless the data collection design includes systematic links, such as raters scoring a subset of the same test-takers as other raters, it is not possible to compare test-takers, raters, and tasks between whom there are no connections. In practice, many operational assessments include these disconnected subsets of assessment components – thereby limiting the comparisons that can be made between test-takers, raters, and tasks. However, when researchers use the Rasch model, they can apply group-anchoring techniques through which they can make comparisons across disconnected subsets. Although researchers and practitioners regularly use group anchoring, there has been limited methodological research related to this technique. In this study, we used simulated data to examine the impact of characteristics of disconnected subsets when group anchoring is used. Our results suggested that the characteristics of disconnected subsets impact the ordering and precision of test-taker estimates, particularly with regard to rating designs and model-data fit within disconnected subsets. We discuss the implications of our findings for research and practice related to rater-mediated assessments.

Keywords: group anchoring; sparse networks; rating designs; performance assessment; Rasch model

[1] *Correspondence concerning this article should be addressed to:* Stefanie A. Wind, PhD, Assistant Professor of Educational Measurement, The University of Alabama, Department of Educational Research Methodology, Box 870231, Tuscaloosa, AL 35487, USA; email: Swind@ua.edu

[2] Department of Educational Psychology, The University of Alabama, USA

In operational administrations of rater-mediated performance assessments, practical constraints often result in incomplete data collection designs, where each rater does not rate each test-taker. Instead, raters often score subsets of test-takers, such that different raters score different test-takers. In order to make meaningful comparisons (e.g., between test-takers and between raters) in these situations, researchers and practitioners can incorporate systematic connections between raters and test-takers into the data collection design, such as raters scoring a subset of the same test-takers as other raters. When these connections are included, researchers and practitioners can use Rasch measurement theory models to obtain estimates of test-taker achievement that are comparable across all raters, and estimates of rater severity that are comparable across all test-takers, even in the presence of large amounts of missing data (Eckes, 2015; Myford & Wolfe, 2000).

In addition to incomplete data collection designs, many operational assessments also lack systematic connections between test-takers, raters, and other facets, such as tasks, prompts, or administrations. Consequently, the design results in *disconnected subsets* (i.e., disjoint subsets; Linacre, 2017) of test-takers and raters between whom there are no connections (see Figure 1 for several examples). Disconnected subsets can occur for a variety of reasons, including a lack of consideration of design issues prior to data collection, purposeful design, practical constraints that prevent the use of connected designs, and failure of the raters to follow the rating design (i.e., judging plan; Sick, 2013). Regardless of their origin, the practical implication of disconnected subsets is that it is not possible to compare test-takers and raters who are nested within different subsets.

Group anchoring

Linacre (2017) proposed a practical strategy for addressing the lack of comparability that results from disconnected subsets. Specifically, when researchers use the Rasch model, they can apply *group anchoring* to facilitate comparisons across disconnected subsets. Group anchoring involves setting the average measure of the groups within one facet, such as raters, test-takers, or tasks, to zero logits. Then, one can estimate locations of the elements within the group-anchored facet relative to the mean of zero logits.

Sick (2013) and Linacre (2017) pointed out several issues that researchers and practitioners should consider when they use group anchoring in their analyses. First, a facet should be group-anchored only when the sample size is sufficient. Specifically, when there is a small sample of observations, extreme values influence the group mean – potentially compromising the interpretation of group anchoring. Linacre discussed the potential influence of extreme values and noted that researchers can choose to exclude extreme values within disconnected subsets during group anchoring (see p. 124). Second, Sick and Linacre recommended that researchers and practitioners use additional information about an assessment to decide which facet should be group-anchored. In particular, when one group-anchors a facet, they are making the assumption that the elements within that facet (e.g., individual test-takers or individual raters) are essentially exchangeable. As an example, Sick pointed out that, if evidence were available that the test-takers within the

Panel A: Complete Blocks (2 disconnected subsets)

Test-Takers	Raters									
	1	2	3	4	5	6	7	8	9	10
1	X	X	X	X	X					
2	X	X	X	X	X					
3	X	X	X	X	X					
4	X	X	X	X	X					
5	X	X	X	X	X					
6						X	X	X	X	X
7						X	X	X	X	X
8						X	X	X	X	X
9						X	X	X	X	X
10						X	X	X	X	X

Panel B: Incomplete Blocks (2 disconnected subsets)

Test-Takers	Raters									
	1	2	3	4	5	6	7	8	9	10
1	X	X								
2		X	X							
3			X	X						
4				X	X					
5	X				X					
6						X	X			
7							X	X		
8								X	X	
9									X	X
10						X				X

Panel C: Complete Blocks (5 disconnected subsets)

Test-Takers	Raters									
	1	2	3	4	5	6	7	8	9	10
1	X	X								
2	X	X								
3			X	X						
4			X	X						
5					X	X				
6					X	X				
7							X	X		
8							X	X		
9									X	X
10									X	X

Panel D: Incomplete Blocks (5 disconnected subsets)

Test-Takers	Raters									
	1	2	3	4	5	6	7	8	9	10
1	X	X								
2		X								
3			X	X						
4				X						
5					X	X				
6						X				
7							X	X		
8								X		
9									X	X
10										X

Figure 1: Illustration of Disconnected Subsets

Note. For each design, raters are shown in the columns and test-takers are shown in the rows. In each cell, an "X" indicates that a rater rated a performance, and a blank cell indicates that a rater did not rate a performance. Disconnected subsets are indicated using borders. For example, in Panel A, there are two disconnected subsets with five raters and five test-takers in each subset. Test-takers 1- 5 cannot be compared to test-takers 6-10 because there are no links between the raters; likewise, Raters 1-5 cannot be compared to Raters 6-10 because there are no links between test-takers. However, test-takers and raters can be compared within each subset.

disconnected subsets had different levels of ability, then a facet besides test-takers, such as raters, should be group-anchored. Finally, researchers and practitioners should clearly state their assumptions related to equivalence when they use group anchoring in order to make the interpretation of the resulting measures transparent.

Because disconnected subsets frequently appear in practical settings, researchers and practitioners have used group anchoring as a post-hoc strategy for making comparisons between components of an assessment procedure that are not connected (discussed further below). However, there is a lack of published studies in which researchers have considered methodological issues related to group anchoring. In particular, researchers have not yet considered the degree to which the following characteristics of disconnected subsets impact the effectiveness of group anchoring: (1) connectivity within disconnected subsets; (2) sample size similarity of disconnected subsets; and (3) model-data fit of the elements in disconnected subsets. In this study, we use simulated data to begin to examine the impact of these characteristics on group anchoring.

Purpose

The purpose of this study is to explore the impacts of connectivity within subsets, sample size, and model-data fit on the relative ordering and precision of test-taker achievement estimates when group anchoring is used in rater-mediated assessments. We use simulated data to address the following research questions:

1. What is the impact of connectivity (i.e., the rating design) within disconnected subsets on the relative ordering and precision of test-taker estimates when group anchoring is used?
2. What is the impact of sample size similarity of disconnected subsets on the relative ordering and precision of test-taker estimates when group anchoring is used?
3. What is the impact of model-data misfit on conclusions on the relative ordering and precision of test-taker estimates when group anchoring is used?

Literature review

In a systematic review of literature in which researchers reported methods for evaluating the quality of ratings in rater-mediated language assessments, Wind and Peterson (2017) observed that most researchers do not provide information about the nature of rating designs, including the number of raters assigned to rate each test-taker or the degree to which there are connections among raters and test-takers. As a result, it is difficult to know the extent to which researchers' data collection designs included disconnected subsets in previous studies, as well as the extent to which they attempted to mitigate disconnected subsets using group anchoring.

Nonetheless, we identified several studies in which the authors reported and discussed disconnected subsets. For example, Zhang and Elder (2011) used group anchoring to calculate estimates of test-taker proficiency that could be compared across disconnected subsets. Specifically, these authors reported the results from a rater-mediated assessment of oral language proficiency in which raters (*n* = 39) rated test-taker performances using a 5-category holistic rating scale. The data collection design resulted in two disconnected subsets, where test taker performances were nested within two approximately equally sized groups of raters (*n* = 20 and *n* = 19). Within the disconnected subsets, all raters scored all performances (i.e., a fully crossed rating design). The authors group-anchored the rater facet to facilitate comparisons between test-takers whose performances had been rated by different groups of raters.

Similarly, Nakatsuhara, Inoue, Berry and Galaczi (2016) reported the results from a study of the comparability of face-to-face and video-conferencing formats for a rater-mediated second-language speaking test. In this study, four raters rated 32 test-takers using a 9-category analytic rating scale with four domains (fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation). One rater rated each test-taker's performance, and each rater rated eight test-takers. As a result, the data collection design resulted in four disconnected subsets made up of one rater and eight test-takers each. These authors group-anchored the rater facet (i.e., anchored the four raters at zero logits) to facilitate comparisons between test-takers who were nested within disconnected subsets, such that the effects of the test format on test-taker achievement could be examined.

Other researchers have applied group-anchoring techniques in more-complex assessment contexts, where additional facets besides test-takers and raters are present. For example, Bonk and Ockey (2003) reported the results from a Many-Facet Rasch model analysis of a second language oral proficiency assessment based on a group discussion task. This study of two years of assessment data included ratings from 20 raters of 2,427 test-takers' performances on a group discussion task. The raters rated test-takers' performances using a 9-category analytic rating scale with five domains (pronunciation, fluency, grammar, vocabulary/content, and communicative skills/strategies). Test-takers responded to one of three prompts. Each test-taker received ratings from two raters in all five domains. Disconnected subsets occurred because each test-taker responded to only one of the three prompts. As a result, the researchers group-anchored the prompt facet (i.e., anchored the three prompts at zero logits) to facilitate comparisons of test-taker achievement between prompts.

In other studies, researchers reported methods for resolving disconnected subsets besides group anchoring. For example, Baird, Hayes, Johnson, and Lamprianou (2013) reported the results from a multi-year analysis of rater-mediated assessments in the areas of geography and psychology. In their analysis of rater effects, these researchers observed that there were disconnected subsets related to the years of the assessment administrations, where different groups of raters and test-takers participated in the assessment in each of the three administrations. As a result, it was not possible to compare test-taker performance across the three years. Rather than using group anchoring to facilitate these comparisons, the authors conducted the analysis separately for each year of assessment data and did not make comparisons between years.

Another approach that researchers have reported using to address disconnected subsets is to include additional observations in the analysis that create connections between otherwise nested subsets. For example, Myford and Wolfe (2000) illustrated the consequences of disconnected subsets on the comparability of test-taker achievement estimates. They demonstrated how researchers make comparisons across observations in disconnected subsets by including as few as one "benchmark" (i.e., common or anchor) performance in the analysis that has been rated by all of the raters. Later, Wind and Jones (2017) examined this technique using simulated data, and found that including a small subset of benchmark performances that all raters rate can facilitate comparisons across otherwise disconnected subsets of test-taker performances, regardless of the sample size, achievement level, or model-data fit within the subset of benchmark performances.

As an alternative to benchmark performances, several researchers have reported using test-taker performances on selected-response items as a method for connecting otherwise disconnected subsets of test-takers in rater-mediated assessments. Specifically, Engelhard and Myford (2003), Wolfe, Myford, Engelhard, and Manalo (2007), and Engelhard and Wind (2013) reported using test-taker responses to multiple-choice items that were designed to measure the same or related constructs as performance assessment tasks to connect disconnected subsets. In each of these studies, test-taker performances on constructed-response tasks were nested within disconnected subsets of raters. However, because all of the test-takers responded to all of the multiple-choice items, the multiple-choice item responses provided a link between the disconnected subsets. Although this technique for connecting disconnected subsets through multiple-choice items may be attractive, Wolfe, et al. (2007) observed that this approach resulted in less accurate comparisons between test-takers compared to using benchmark performances.

Methods

We used simulated data to explore the research questions for this study because this approach allowed us to manipulate selected characteristics of disconnected subsets. In order to provide a focused methodological investigation, we used a two-facet assessment design (test-takers and raters) with disconnected subsets. However, other assessment systems include additional facets, such as tasks, prompts, scoring centers, or administrations, and it is also possible to apply group-anchoring techniques to data collected under these designs. Therefore, our simulation conditions should not be viewed as an exhaustive representation of data collection designs in which group anchoring may be used. Rather, our simulation study provides initial insight into the consequences of group anchoring using a relatively simple design.

Simulation procedure

Table 1 lists the variables that we held constant and the variables that we manipulated in our simulation study. Using the specifications in Table 1, we generated 100 datasets (replications) with characteristics defined by each of the 2 × 2 × 2 × 3 × 2 = 48 possible combinations of the levels of the manipulated variables, for a total of 4,800 unique datasets. We generated the ratings using base programming in the *R* statistical software program (R Core Team, 2018).

Table 1:
Summary of the Variables included in the Simulation Study

Variables		**Levels**
Variables held constant	Number of test-takers	25* Number of raters
	Generating test-taker achievement parameters	$\theta \sim N(0,1)$
	Generating rater severity parameters	$\lambda \sim N(0,1)$
	Rating scale length	4 categories using the Andrich Rating Scale Model (RSM) with three thresholds (τ_1, τ_2, τ_3), where τ_1 was randomly selected from U(0, -3), τ_3 was randomly selected from U(0, 3), and τ_2 was selected such that $\tau_{1+} \tau_2 + \tau_3 = 0$.
Variables manipulated	Rater sample size	20; 50
	Rating Design within Disconnected Subsets	Complete blocks; Incomplete blocks with systematic links
	Number of disconnected subsets	Rater N/5; Rater N/10
	Proportion of raters modeled to exhibit model-data misfit	5% of all raters; 20% of raters in half of the subsets & 5% in other half of subsets; 20% in all subsets
	Sample size balance across disconnected subsets	Balanced (all subsets have the same number of test-takers); Unbalanced (half of the subsets have twice as many test-takers as the other half)

Simulating incomplete ratings

In order to explore the effects of disconnected subsets, we started by generating complete ratings based on the specified sample size and other generating parameter characteristics (described below), where every rater scored every test-taker (i.e., there was no missing data). Then, we created disconnected subsets by removing from the generated datasets certain observations for certain raters or test-takers in order to create disconnected subsets, with characteristics specified by the simulation condition. In the conditions with incomplete ratings, two raters rated each test-taker, such that the rating design within each subset resembled Figure 1, Panel B.

Variables held constant

As shown in Table 1, we held four variables constant in our simulation design. First, we used the same number of test-takers in all conditions. Our test-taker sample size was equal to 25 times the number of raters included in the simulation condition. This ratio of 25 test-takers to one rater reflects current practice in educational performance assessments, as well as the sample sizes that researchers have reported in real-data studies of rater-mediated assessments (e.g., Duckor, Castellano, Téllez, Wihardini, & Wilson, 2014; Wilson & Hoskens, 2001). Second, we used the same distribution of generating test-taker achievement parameters (i.e., theta estimates) for all of the simulation conditions. Following the procedures that researchers have used in previous simulations of rater-mediated performance assessments (e.g., Marais & Andrich, 2011; Meyer & Hailey, 2012; Wolfe, Jiao, & Song, 2014), we generated test-taker achievement parameters and rater severity parameters from a normal distribution with a mean of zero logits and a standard deviation of one logit for all of the simulation conditions. Finally, we used the same rating scale length in all of the simulation conditions. Specifically, we modeled a 4-category rating scale in order to match current practice in many large-scale educational performance assessments (e.g., Commonwealth of Virginia, Department of Education, 2012), as well as recent simulation studies of rating scale data (e.g., Meyer & Hailey, 2012).

Variables manipulated

To explore the effects of different characteristics of disconnected subsets, we manipulated five variables in our simulation design. First, we used two rater sample sizes to represent performance assessments of different sizes: 20 raters or 50 raters. Second, we modeled two different types of rating designs within the disconnected subsets in our simulated ratings: complete blocks or incomplete blocks with systematic links (see Figure 1); these designs reflect rating designs that are appropriate for analysis with Rasch measurement theory. Furthermore, researchers have used these types of incomplete rating designs in previous Rasch model analyses of rater-mediated assessments (Engelhard, 1997; Hombo, Donoghue, & Thayer, 2001). Third, we included two different numbers of disconnected subsets in our simulation study determined using the number of raters. Specifically, we included disconnected subsets equal to the rater sample size divided by five (N_{raters} = 20: 4 disconnected subsets; N_{raters} = 50: 10 disconnected subsets) or the rater sample size divided by ten (N_{raters} = 20: 2 disconnected subsets; N_{raters} = 50: 5 disconnected subsets).

To explore the influence of model-data misfit on group anchoring, we modeled three different proportions of raters to exhibit model-data misfit. First, as a baseline condition, we modeled 5% of the raters to exhibit model-data misfit, where we randomly selected raters from all of the subsets to exhibit misfit. Second, we simulated a higher magnitude of misfit by modeling 20% of the raters to exhibit misfit in half of the subsets. In the other half of the subsets, we randomly selected 5% of the raters to exhibit misfit. Third, we simulated misfit in all of the disconnected subsets by randomly selecting 20% of the raters from each subset and modeling them to exhibit misfit. We modeled rater misfit by drawing the generating rater slope parameters from $U \sim [0.3, 0.7]$. For the raters who fit the Rasch model, we fixed the generating rater slope parameters to the usual Rasch value (1.0). To simulate misfit, we multiplied the slope parameters by the difference between the test-taker achievement parameters and rater severity parameters (θ - λ) during our simulation procedure. The result of this process was that, when we analyzed the ratings using the RSM, the raters with slopes besides 1.0 misfit the model.

Finally, we considered the influence of the similarity of sample sizes across disconnected subsets. Specifically, we designed the simulation to produce either *balanced* or *unbalanced* sample sizes in different subsets. For the conditions with balanced subsets, we modeled all of the disconnected subsets to include the same number of test-takers and raters. On the other hand, for the conditions with unbalanced subsets, we included twice as many test-takers in half of the subsets as the other half of the subsets. We included the same number of raters in all of the subsets.

Data analysis

To explore the effects of different characteristics of disconnected subsets on the values and precision of test-taker estimates, we conducted a three-step data analysis procedure. First, we used the Facets software program (Linacre, 2015) to estimate test-taker locations (i.e., achievement estimates for each performance) based on the polytomous Rasch model (Andrich, 1978):

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

where θ is the test-taker estimate on the logit scale, λ is the rater severity estimate on the logit scale, and τ is the threshold where there is an equal probability for a rating in category k and category k – 1. We used a Rasch model because this approach is used most often with the group anchoring technique (Linacre, 2017). In order to obtain test-taker estimates in the presence of disconnected subsets, we applied group anchoring by setting the mean of the rater estimates (λ) to zero logits within each subset. We provided the Facets program syntax in Appendix A, which shows our procedure for group anchoring.

After we calculated the test-taker estimates from each of the generated datasets, we compared the estimates to the generating parameters that we used to simulate the ratings. Specifically, we used the Pearson product-moment correlation (r) for this comparison. Following Corey, Dunlap, and Burke's (1998) recommendations for averaging correlations, we used the following procedure to summarize the correlations for each of the simulation conditions: (1) calculate r between each set of 100 corresponding generating and estimated test-taker parameters for each simulation condition, (2) use the psych package for *R* (Revelle, 2016) to convert each r value to a standardized correlation using Fisher's z transformation, (3) calculate the average z value over all of the replications of each simulation condition, (4) convert the average z value back to its corresponding r value (r_z) to facilitate interpretation.

In addition to the relative ordering of test-taker estimates, we were also interested in the effects of different characteristics of disconnected subsets and group anchoring on the *precision* of test-taker estimates. Precision of estimates from IRT models can be examined in several ways. In this study, we used the standard error (*SE*) of the test-taker estimates as an indicator of measurement precision, where smaller standard errors reflect more precise estimates, and larger standard errors reflect less precise estimates. We used the Facets program to obtain the *SE* for each test-taker estimate.

Finally, we followed Harwell, Stone, Hsu, and Kirisci's (1996) recommendation that researchers use ANOVA models to summarize the impact of simulation design factors when simulation studies are based on factorial designs. Specifically, we used the average r_z and *SE* as dependent variables in two separate ANOVA models. We used the five manipulated variables in the simulation design (rater sample size, number of disconnected subsets, proportion of raters modeled to exhibit misfit, rating design within subsets, and balance of sample sizes across subsets) as independent variables. We used model building to retain only statistically significant effects based on $\alpha = 0.01$.

Results

Before we compared the consistency of test-taker ordering and the precision of test-taker estimates, we examined the results from the polytomous Rasch model to ensure that the simulation procedure produced datasets that matched our specifications. When we checked the average test-taker estimates and rater estimates, we observed that the distributions of these values were close to the specified distributions ($M = 0$, $SD = 1$). Furthermore, we checked the Rasch Infit and Outfit mean square error (*MSE*) and standardized model-data fit statistics for test-takers, and we found that, on average, the values of the unstandardized and standardized versions of these statistics were near the values that previous researchers have established as expected when there is acceptable fit to the model (*MSE* Infit and Outfit around 1.00, standardized Infit and Outfit between +2 and -2; Smith, 2004; Wu & Adams, 2013). We also checked the fit statistics for the raters who we did *not* model to misfit the Rasch model, and the average fit statistics for these raters were also within the previously established range for acceptable fit. Specifically, the average *MSE* Infit and Outfit statistics ranged from 0.82 to 1.02 and the average standardized Infit and Outfit statistics ranged from -0.90 to 0.17 for these raters over all of the simulation conditions. These slightly low average values of model-data fit statistics (described as “overfit” by some researchers) reflect the presence of the misfitting raters in our simulated datasets. Finally, we checked the fit statistics for the raters who we modeled to exhibit misfit and observed that the average fit statistics for these raters were outside the range of expected values when data fit the Rasch model. Specifically, the average *MSE* Infit and Outfit statistics ranged from 1.62 to 3.18 and the average standardized Infit and Outfit statistics ranged from -2.30 to 5.23 for these raters over all of the simulation conditions.

Relative ordering of test-taker estimates

Table 2 presents a summary of the results of the correlation analysis between the generating test-taker parameters and the test-taker estimates. For each condition, we report the average standardized correlation coefficient (based on Fisher’s transformation of Pearson’s r) over 100 replications of the condition transformed back to the Pearson r scale (r_z). Higher values of r_z indicate a closer correspondence between the relative ordering of the generating test-taker parameters and the test-taker estimates, indicating that the ordering of test-takers remained consistent when group anchoring was used to resolve disconnected subsets. We chose to present the correlations on the r_z scale, rather than reporting disattenuated correlation coefficients for two reasons. First, we wanted to highlight the impact of measurement error on the correspondence between the generating parameters and estimates of test-taker achievement. Second, we followed Muchinsky’s (1996) advice that researchers should use un-corrected correlation coefficients in statistical hypothesis tests, such as ANOVA models. Because we used ANOVA to summarize the results from our correlation analysis, the un-corrected values were more appropriate. For all of the simulation conditions, the standard deviations of r_z were 0.04 or less.

Results from the ANOVA where r_z was the dependent variable indicated that rater sample size did not have a statistically significant effect on mean r_z. Furthermore, there were no statistically significant interactions between rater sample size and any of the other manipulated characteristics of the disconnected subsets. Similarly, the ANOVA results did not indicate a statistically significant effect for the balance of sample sizes across disconnected subsets. We summarized the results from our final ANOVA model for r_z in Table 3. In terms of the effect of the number of disconnected subsets, the ANOVA results indicated a non-significant main effect for this factor ($F(1, 42) = 1.25$, $p = 0.27$, $\eta^2 = 0.03$). However, the interaction between the number of disconnected subsets and the type of rating design within subsets was statistically significant ($F(1, 42) = 7.25$, $p < 0.01$, $\eta^2 = 0.15$).

Table 2:
Average Correlations (r_z) between Generating Theta Parameters and Theta Estimates across 100 Replications of Simulation Conditions

Rater *N*	**Number of Disconnected Subsets**	**Proportion of raters modeled to exhibit model-data misfit**	**Complete Rating Designs**		**Incomplete Rating Designs**	
			Balanced Sample Size	**Unbalanced Sample Size**	**Balanced Sample Size**	**Unbalanced Sample Size**
20	2 (Rater *N*/10)	5% in all subsets	0.90	0.71	0.35	0.34
		20% of raters in half of the subsets; 5% in half of the subsets	0.71	0.71	0.34	0.32
		20% in all subsets	0.70	0.70	0.33	0.31
	4 (Rater *N*/5)	5% in all subsets	0.80	0.68	0.56	0.55
		20% of raters in half of the subsets; 5% in half of the subsets	0.66	0.65	0.55	0.54
		20% in all subsets	0.66	0.65	0.55	0.54
50	2*	5% in all subsets	0.95	0.74	0.58	0.58
		20% of raters in half of the subsets; 5% in half of the subsets	0.74	0.73	0.57	0.57
		20% in all subsets	0.73	0.73	0.55	0.56
	5 (Rater *N*/10)	5% in all subsets	0.67	0.56	0.41	0.37
		20% of raters in half of the subsets; 5% in half of the subsets	0.59	0.57	0.42	0.37
		20% in all subsets	0.57	0.55	0.40	0.36
	10 (Rater *N*/5)	5% in all subsets	0.84	0.69	0.55	0.54
		20% of raters in half of the subsets; 5% in half of the subsets	0.69	0.69	0.54	0.53
		20% in all subsets	0.68	0.68	0.51	0.51

* *Note.* We added additional simulation conditions with two subsets for *N* Raters = 50 based on our observation of the relatively weak correlations for the two-subset conditions when *N* Raters = 20. However, as the additional conditions were not part of our simulation design, we did not include them in our ANOVA models.

Table 3:
ANOVA Model I: Dependent Variable = Correlation between Generating Theta Parameters and Theta Estimates (r_z)

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Number of Disconnected Subsets	0.01	1	0.01	1.25	0.27	0.03
Proportion of Raters modeled to Misfit	0.02	2	0.01	1.70	0.20	0.08
Rating Design	0.64	1	0.64	93.08	< 0.01	0.69
Number of Subsets * Rating Design	0.05	1	0.05	7.25	< 0.01	0.15
Error	0.29	42	0.01			

Specifically, the average correlations for the complete subset conditions were higher when there were fewer disconnected subsets (Rater *N*/10: $0.68 \leq M_{rz} \leq 0.90$) compared to more disconnected subsets (Rater *N*/5: $0.55 \leq M_{rz} \leq 0.80$). However, the number of disconnected subsets appeared to have the opposite effect when the disconnected subsets contained incomplete ratings – in these cases, the average correlations were higher when there were more disconnected subsets (Rater *N*/5: $0.37 \leq M_{rz} \leq 0.56$) compared to fewer disconnected subsets (Rater *N*/10: $0.31 \leq M_{rz} \leq 0.55$).

In terms of the effect of the proportion of raters modeled to exhibit misfit, the results in Table 2 indicate that, regardless of rater sample size, the correlations between generating parameters and test-taker estimates were generally higher when 5% of the raters were modeled to exhibit model-data misfit ($0.34 \leq M_{rz} \leq 0.90$) compared to conditions where 20% of the raters were modeled to exhibit model-data misfit in half of the subsets ($0.32 \leq M_{rz} \leq 0.71$) and the conditions where 20% of the raters were modeled to exhibit misfit in all of the subsets ($0.31 \leq M_{rz} \leq 0.70$). However, the results from the ANOVA suggested that the proportion of raters modeled to exhibit misfit did not have a statistically significant effect on the average correlation coefficient (see Table 3).

Figure 2 illustrates the consequences of disconnected subsets on the correspondence between generating test-taker parameters (theta parameters) and test-taker estimates (theta estimates) for the simulation condition in which we observed the lowest average value of r_Z. Specifically, we randomly selected one replication from the simulation conditions with Rater $N = 20$, 20% of raters modeled to exhibit misfit, incomplete rating designs within subsets, and un-balanced sample sizes across subsets (*Mean* r_Z for this condition = 0.31). In the figure, we used different plotting symbols to represent students in the two disconnected subsets. Inspection of this plot highlights the general lack of correspondence between the theta parameters and theta estimates for both subgroups. The placement of individual estimates below or above the black identity line illustrates the consequences of the characteristics of the simulation condition for individual test-takers, where some test-takers have higher estimates than their true (generating) parameter values, and other test-takers have lower estimates than their parameter values.

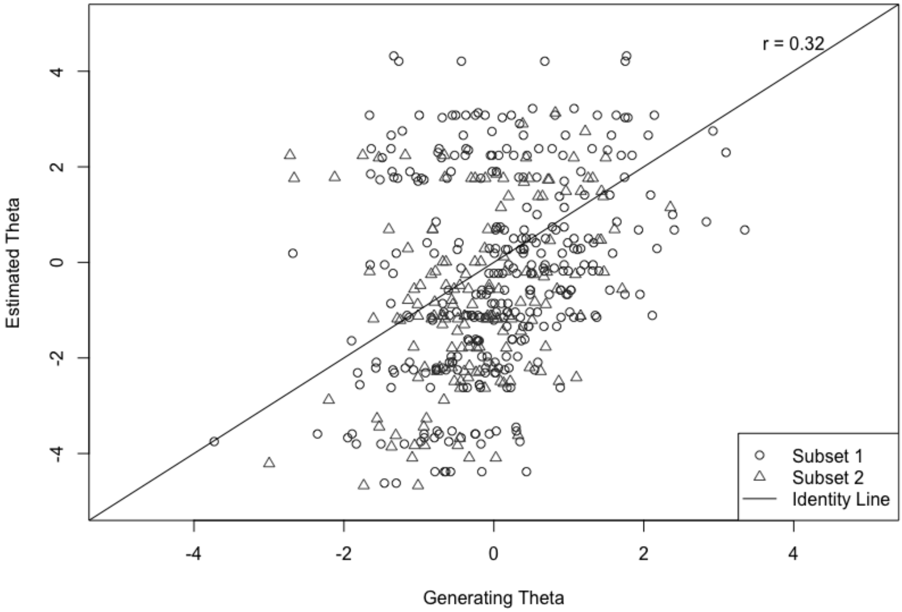

Figure 2:
Correspondence between Generating Thetas and Estimated Thetas within a Randomly Selected Replication of the Rater N = 20, Two-Subset, Incomplete Design, Unbalanced Sample Size Condition with 20% of Raters Modeled to Exhibit Misfit

After we observed the relatively weak correlations in the incomplete rating design conditions for Rater *N* = 20 where there were two disconnected subsets, we added additional simulation conditions for Rater *N* = 50 with two disconnected subsets. Because two-subset designs are relatively common in practice (e.g., assessments that involve two administrations, two days of scoring, two test centers, etc.), including these additional conditions allowed us to understand whether the relatively poor results from the two-subset conditions under Rater *N* = 20 generalize to larger sample sizes. However, because these conditions were not part of our original design, and in order to avoid a lack of balance in conditions, we did not include results from these conditions in our ANOVA models. We observed generally similar patterns between the complete and incomplete rating designs in these conditions, where the average values of r_z were notably higher for the complete rating designs ($0.73 \leq r_z \leq 0.95$) compared to the incomplete rating designs ($0.55 \leq r_z \leq 0.58$). We also saw similar effects with regard to the effect of rater misfit, where average values of r_z decreased when we modeled more raters to exhibit model-data misfit.

Precision of test-taker estimates

Table 4 presents the mean of the standard errors (*SE*) of the test-taker estimates over 100 replications of each simulation condition. For all of the simulation conditions, the standard deviations of the *SE*s were 0.05 or less. In this table, higher values indicate less precision, and lower values indicate more precise estimates. Results from the ANOVA with the average test-taker *SE* as the dependent variable indicated that, similar to r_z, rater sample size did not have a statistically significant effect on mean r_z. Furthermore, there were no statistically significant interactions between rater sample size and any of the other manipulated characteristics of the disconnected subsets. Similarly, the ANOVA results did not indicate a statistically significant effect for the balance of sample sizes across disconnected subsets. We summarized the results from our final ANOVA model for SE_θ in Table 5.

In terms of the effect of the number of disconnected subsets, the results suggest that the average *SE* was higher, suggesting less precision, in the simulation conditions with more disconnected subsets (Rater *N*/5) compared to the simulation conditions with fewer disconnected subsets (Rater *N*/10). The statistically significant main effect for the number of disconnected subsets ($F(1, 42) = 722.49$, $p < 0.01$, $\eta^2 = 0.95$) reflects this result. Furthermore, there was a statistically significant main effect for the type of rating design within subsets ($F(1, 42) = 17{,}652.81$, $p < 0.01$, $\eta^2 = 0.998$). In particular, regardless of all of the other characteristics of the disconnected subsets, the average *SE* in conditions with incomplete subsets was nearly twice the size of the corresponding average *SE* in conditions with complete subsets. For example, in the conditions based on incomplete rating designs, the average *SE* ranged from 1.18 to 1.34. The average *SE* was notably lower in the conditions based on complete rating designs, ranging from 0.40 to 0.70.

However, there was also a statistically significant interaction between the number of disconnected subsets and the type of rating design within subsets ($F(1, 42) = 276.14$, $p < 0.01$, $\eta^2 = 0.87$). Specifically, the effect of the number of subsets on the average *SE* appeared to depend on the type of rating design that was used within subsets. For the conditions based on Rater *N*/5 disconnected subsets and *complete* rating designs, the average *SE*s range from 0.62 to 0.70, and the average *SE*s for the Rater *N*/5 disconnected subsets conditions with *incomplete* designs range from 1.23 to 1.34. We observed a similar pattern for the Rater *N*/10 disconnected subset conditions: When there were complete ratings within the disconnected subsets, the average *SE* ranged from 0.41 to 0.44, and when there were incomplete ratings within the disconnected subsets, the average *SE* ranged from 1.18 to 1.28.

The ANOVA results also revealed a statistically significant main effect for the proportion of raters who were modeled to exhibit model-data misfit ($F(2, 42) = 22.29$, $p < 0.01$, $\eta^2 = 0.52$). The results in Table 4 indicate that the average *SE*s were generally higher when we modeled fewer raters to exhibit model-data misfit compared to the conditions where we modeled more raters to exhibit model-data misfit. However, the differences in the average *SE*s across the three levels of model-data misfit were relatively small within types of rating designs and numbers of disconnected subsets. For example, in the Rater

Table 4:
Average Standard Errors of Test-Taker Estimates across 100 Replications of Simulation Conditions

Rater N	Number of Disconnected Subsets	Proportion of raters modeled to exhibit model-data misfit	Complete Rating Designs		Incomplete Rating Designs	
			Balanced Sample Size	Unbalanced Sample Size	Balanced Sample Size	Unbalanced Sample Size
20	2 (Rater N/10)	5% in all subsets	0.44	0.45	1.27	1.28
		20% of raters in half of the subsets; 5% in half of the subsets	0.42	0.43	1.25	1.25
		20% in all subsets	0.41	0.41	1.22	1.24
	4 (Rater N/5)	5% in all subsets	0.66	0.70	1.29	1.29
		20% of raters in half of the subsets; 5% in half of the subsets	0.65	0.65	1.28	1.28
		20% in all subsets	0.62	0.63	1.23	1.25
50	2*	5% in all subsets	0.26	0.28	1.40	1.41
		20% of raters in half of the subsets; 5% in half of the subsets	0.26	0.26	1.35	1.34
		20% in all subsets	0.25	0.25	1.32	1.33
	5 (Rater N/10)	5% in all subsets	0.44	0.44	1.26	1.23
		20% of raters in half of the subsets; 5% in half of the subsets	0.42	0.41	1.25	1.21
		20% in all subsets	0.41	0.40	1.20	1.18
	10 (Rater N/5)	5% in all subsets	0.68	0.70	1.32	1.34
		20% of raters in half of the subsets; 5% in half of the subsets	0.67	0.67	1.31	1.33
		20% in all subsets	0.64	0.65	1.28	1.31

* Note. We added additional simulation conditions with two subsets for N Raters = 50 based on our observation of the relatively weak correlations for the two-subset conditions when N Raters = 20. However, as the additional conditions were not part of our simulation design, we did not include them in our ANOVA models.

Table 5:
ANOVA Model II: Dependent Variable = Standard Error of Theta Estimates

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***	**Partial Eta Squared**
Number of Disjoint Subsets	0.26	1	0.26	722.49	< 0.01	0.95
Proportion of Raters modeled to Misfit	0.02	2	0.01	22.29	< 0.01	0.52
Rating Design	6.27	1	6.27	17652.81	< 0.01	0.998
Number of Subsets * Rating Design	0.10	1	0.10	276.14	< 0.01	0.87
Error	0.02	42	0.00			

N/5 conditions based on complete rating designs, the average *SE* was similar among the conditions where 5% of the raters were modeled to exhibit model-data misfit ($0.66 \leq M_{SE} \leq 0.70$), the conditions were 20% of the raters were modeled to exhibit misfit in half of the subsets ($0.65 \leq M_{SE} \leq 0.67$), and the conditions were 20% of raters in all of the subsets were modeled to exhibit misfit ($0.62 \leq M_{SE} \leq 0.65$).

We also included average *SE*s for the additional simulation conditions with Rater $N = 50$ and two disconnected subsets. The results from these conditions followed the general patterns that we observed in the other conditions. Specifically, in these conditions, the average *SE*s were notably lower when the disconnected subsets included complete rating designs compared to the incomplete rating design conditions. Furthermore, the average *SE*s were slightly lower (indicating more precision) when more raters were modeled to exhibit model-data misfit.

Discussion

The purpose of this study was to explore the impacts of different characteristics of disconnected subsets on the relative ordering and precision of test-taker estimates when group anchoring is used. We simulated ratings with disconnected subsets of raters and test-takers that varied in terms of five characteristics: rater sample size, number of disconnected subsets, proportion of raters modeled to exhibit misfit, rating design within disconnected subsets, and the balance of test-taker sample sizes across disconnected subsets. We analyzed the ratings using the polytomous Rasch model, with the rater facet group-anchored. Overall, our results suggested that differences in the characteristics of disconnected subsets impacted the ordering and precision of test-taker estimates. In this section, we return to each of our research questions and provide tentative conclusions based on the results from our simulation study.

The first research question focused on the impact of the type of rating design within disconnected subsets on the relative ordering and precision of test-taker estimates when group anchoring is used. To address this research question, we simulated disconnected subsets that included either *complete ratings* where every rater rated every test-taker, or *incomplete ratings*, where there were systematic links between raters and test-takers, but every rater did not rate every test-taker. Our results indicated that the type of rating design had a notable effect on the correlation between test-taker estimates and generating test-taker parameters, as well as the *SE* of test-taker estimates. Incomplete rating designs were associated with less consistency in relative test-taker ordering with the generating parameters, as well as less precision in test-taker estimates. In other words, when there are fewer links between raters and test-takers within disconnected subsets, the relative ordering of test-takers is more likely to diverge from the ordering that would have been observed without disconnected subsets compared to when there are complete ratings within disconnected subsets. Likewise, when there are fewer links between raters and test-takers within disconnected subsets, there is likely to be more error associated with the achievement estimates for test-takers compared to when complete ratings are used within disconnected subsets.

An important result related to the rating design specification was the impact of complete or incomplete designs on average *SE*s for test-takers. The relatively large average test-taker *SE*s for the incomplete rating designs were related to the large amount of missing data in these conditions, where we only had two observations for each test-taker. Nonetheless, the magnitude of the difference in average *SE*s between the conditions with complete and incomplete rating designs should remind researchers and practitioners of the effects of large proportions of missing data on the precision of test-taker estimates in general, and in combination with disconnected subsets in particular. Although it is possible to estimate test-taker achievement without a complete rating design, sparse connections are not without consequence to the precision of test-taker estimates.

The second research question focused on the impact of sample size similarity across disconnected subsets (i.e., balanced samples) on the relative ordering and precision of test-taker estimates when group anchoring is used. In order to address this research question, we simulated disconnected subsets that included either *balanced* sample sizes across disconnected subsets, where every subset included the same number of test-takers, or *unbalanced* sample sizes across disconnected subsets, where there were differences in the number of test-takers across disconnected subsets. Our results indicated that the similarly of the number of test-takers between different subsets did not have a substantial effect on either the relative ordering of test-taker estimates or the precision of the estimates.

Our third research question focused on the impact of rater misfit on the relative ordering and precision of test-taker estimates when group anchoring is used. In order to address this research question, we modeled three proportions of raters to exhibit misfit: 5% of all raters, 20% of raters in half of the subsets and 5% in the other half, or 20% of raters in all of the subsets. Our results indicated that the proportion of raters modeled to exhibit misfit did not have a significant effect on the consistency of test-taker estimates with the generating parameters. However, the proportion of raters modeled to exhibit misfit had a significant effect on the precision of test-taker estimates. Interestingly, when more raters were modeled to exhibit misfit, the test-taker estimates were more precise compared to the simulation conditions where fewer raters were modeled to exhibit misfit.

The finding of more-precise estimates with higher proportions of raters modeled to exhibit misfit is interesting and warrants further consideration. After we observed this result, we examined the estimated rating scale category thresholds in the conditions with different proportions of raters modeled to exhibit misfit. In this analysis, we observed a slightly wider spread of the estimated thresholds, indicating a more frequent use of the middle rating scale categories, in the conditions where we modeled more raters to exhibit model-data misfit. However, because we used the RS model, it was not possible to examine threshold estimates separately for raters who we modeled to exhibit misfit and raters who we did not model to exhibit misfit. In future studies, researchers should examine the precision of test-taker estimates calculated using the Partial Credit model (Masters, 1982) in the context of group-anchoring designs. It is also possible that the larger average *SE*s that we observed in the simulation conditions in which we modeled 5% of raters to exhibit misfit and the conditions in which we modeled 20% of raters modeled to exhibit in half of the subsets and 5% in the other half of the subsets were related to *differences* in rater fit across the subsets. For example, when the baseline level of rater misfit was specified, we randomly selected 5% of the raters to exhibit misfit; these raters were not necessarily evenly distributed across the subsets. Similarly, the conditions where we modeled 20% of raters to exhibit misfit in half of the subsets resulted in larger discrepancies in rater fit across subsets compared to the conditions where we modeled the same amount of misfit in all of the subsets. Second, it is interesting to note that our finding of smaller average *SE*s when more raters were modeled to exhibit misfit is similar to a finding reported by Wind and Jones (2017). Specifically, these authors used simulated data to examine the influence of the characteristics, including model-data fit, of test-takers used to establish links across otherwise disconnected subsets in performance assessments, on the consistency and precision of parameter estimates. Their results suggested that common test-takers that were modeled to exhibit misfit resulted in more consistent and precise estimates compared to common test-takers that were modeled to exhibit acceptable model-data fit. As Wind and Jones noted, the somewhat stabilizing effects of model-data misfit in both cases could be related to the presence of more “noise” (i.e., variation), which provides more statistical information about individual test-takers or raters.

Implications for practice

Rating designs play an important role in the interpretation of both raw-score (i.e., observed) ratings and estimates of test-taker achievement based on latent trait models. However, most researchers who report the results from studies of rater-mediated performance assessments do not include details about rating designs (Wind & Peterson, 2017). Likewise, many operational assessments do not make information about rating designs available. As a result, it is not possible to determine whether there is sufficient connectivity to warrant comparisons across raters and test-takers.

Ideally, the best solution to overcome challenges related to disconnected subsets is to avoid them altogether. Accordingly, we encourage researchers and practitioners to begin data analyses as soon as possible after data collection begins in order to identify and resolve shortcomings in the rating design (as implemented), such as disconnected subsets. However, if it is not possible to establish connectivity during scoring for a rater-mediated assessment, group-anchoring techniques based on the Rasch model provide a solution to resolve disconnected subsets. Accordingly, several researchers have recommended this approach to researchers and practitioners whose data include disconnected subsets (e.g., Linacre, 2017; Sick, 2013).

Using a simulation study, we observed that the characteristics of disconnected subsets matter with regard to both the stability of performance ordering and the precision of performance estimates. In particular, the rating design within disconnected subsets and the proportion of raters who were modeled to exhibit misfit had substantial effects on the precision and stability of performance estimates. This finding suggests that if group anchoring is used to resolve disconnected subsets in data from a rater-mediated assessment, the degree to which raters are connected and exhibit misfit within the disconnected subsets can potentially influence the results of comparisons between test-takers based on their achievement estimates.

In most situations in which disconnected subsets occur, it is not possible for researchers and practitioners to specify ahead of time what types of rating design should be used within disconnected subsets; if this were possible, one would probably avoid disconnected subsets altogether. Furthermore, even with rater training and monitoring, it is often not possible to ensure that raters exhibit acceptable levels of model-data fit during operational scoring. Therefore, the results from this study should not be interpreted as recommendations for the design of data collection systems with disconnected subsets. Rather, our findings should alert researchers and practitioners to the potential impact of characteristics of disconnected subsets related to rating designs and model-data fit on performance estimates when group anchoring is used. In particular, this information can help researchers and practitioners appropriately qualify their claims when comparing and classifying test-takers based on estimates obtained using group anchoring.

Nonetheless, the results from our study do provide some tentative guidance for the design of rater-mediated assessments. When researchers and practitioners know ahead of time that disconnected subsets are unavoidable, such as in assessment contexts with two separate scoring centers or assessments with two separate administrations, they can use the results from this study to make decisions about characteristics that can be controlled, such as rater sample size, rating designs within subsets, and the number of subsets. In particular, the average correlation coefficients in Table 2 reveal important differences in the accuracy of test-taker achievement estimates under different conditions. For example, if a two-subset design with incomplete ratings is necessary, it may be desirable to include 50 or more raters, as these designs were associated with notably higher average correlations between generating parameters and test-taker estimates ($0.55 \leq r_z \leq 0.58$) than the corresponding conditions with 20 raters ($0.31 \leq r_z \leq 0.35$). Researchers and practitioners should make such decisions in light of consideration of the unique assessment context, including the purpose of the assessment, the intended use of the assessment results, and the anticipated consequences of the assessment.

Limitations and directions for future research

Our study has several limitations that researchers and practitioners should consider before generalizing the results to contexts beyond the simulation study. First, we used a limited set of simulation conditions so that we could focus our analyses on specific characteristics of disconnected subsets. Particularly, we focused on a simple assessment context with only two facets: Test-takers and raters. In practice, performance assessments may include different characteristics, such as different types of rating designs within disconnected subsets (e.g., anchor or spiral designs), different rating scale lengths, different sample sizes of test-takers and raters, or multiple items or prompts. Similarly, raters may exhibit effects other than misfit, such as central tendency or severity/leniency effects. In addition, other types of model-data misfit, including person misfit or task misfit, may be present. Further, it is important to note that our simulation procedure resulted in good targeting between rater severity and test-taker achievement. In operational assessment contexts, there may be less alignment between raters and students. In future studies, researchers should consider the effects of additional characteristics beyond those examined here, including situations with more than two facets, additional types of misfit, and different levels of targeting, in order to determine the degree to which our findings extend to a wider range of performance assessment contexts.

It is also important to note that our study was based on group-anchoring the rater facet only, and we did not explore the implications of group-anchoring the test-taker facet. As we discussed earlier in the manuscript, when one group-anchors a facet, they are making the assumption that the elements within that facet (e.g., individual test-takers or individual raters) are essentially exchangeable. Accordingly, researchers usually do not group-anchor the facet that they are directly investigating (i.e., the object of measurement); in many cases, this is the test-taker. However, there may be cases, such as in rater-effect analyses, where differences between raters are the focus. In these situations, researchers could group-anchor test-takers, assuming that they are essentially exchangeable, in order to identify differences among raters. In future studies, researchers should examine the extent to which characteristics of disconnected subsets impact the accuracy and precision of rater estimates when test-takers are group-anchored.

References

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 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 generaliza-*

bility theory, Rasch modelling and multilevel modelling (No. Ofqual/13/5261). Coventry, United Kingdom: Office of Qualifications and Examinations Regulation.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Commonwealth of Virginia, Department of Education. (2012). *Virginia Standards of Learning Assessments Test Blueprint: End of Course Writing*. Richmond, VA. Retrieved from http://www.doe.virginia.gov/testing/sol/blueprints/english_blueprints/2010/2010_blueprint_eoc_writing.pdf

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations. *The Journal of General Psychology*, *125*(3), 245–261. https://doi.org/10.1080/00221309809595548

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the Internal Structure Evidence for the Performance Assessment for California Teachers: A Validation Study of the Elementary Literacy Teaching Event for Tier I Teacher Licensure. *Journal of Teacher Education*, *65*(5), 402–420. https://doi.org/10.1177/0022487114542517

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

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

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*. New York: College Entrance Examination Board.

Engelhard, G., & Wind, S. A. (2013). *Rating Quality Studies Using Rasch Measurement Theory* (Research Report No. 2013–3). New York, NY: The College Board.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

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.

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

Linacre, J. M. (2017). A user's guide to FACETS: Rasch-model computer programs. winsteps.com. Retrieved from http://www.winsteps.com/manuals.htm

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

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*(3), 248–258.

Muchinsky, P. M. (1996). The Correction for Attenuation. *Educational and Psychological Measurement*, *56*(1), 63–75. https://doi.org/10.1177/0013164496056001004

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

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2016). *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery: A preliminary comparison of test-taker and examiner behaviour* (No. IELTS Partnership Research Papers 1). IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from https://www.ielts.org/teaching-and-research/research-reports/ielts-partnership-research-paper-1

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/

Revelle, W. (2016). psych: Procedures for Personality and Psychological Research (Version 1.6.9). Evanston, Illinois, USA: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Sick, J. (2013). Judging plans and disjoint subsets. *Shiken Research Bulletin*, *17*(1), 27–32.

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.

Wilson, M., & Hoskens, M. (2001). The Rater Bundle Model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306.

Wind, S. A., & Jones, E. (2017). The Stabilizing Influences of Linking Set Size and Model – Data Fit in Sparse Rater-Mediated Assessment Networks. *Educational and Psychological Measurement*, 001316441770373. https://doi.org/10.1177/0013164417703733

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 026553221668699. https://doi.org/10.1177/0265532216686999

Wolfe, E. W., Jiao, H., & Song, T. (2014). A Family of Rater Accuracy Models. *Journal of Applied Measurement*, *16*(2), 153–160.

Wolfe, E. W., Myford, C. M., Engelhard, G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition Examination using benchmark essays* (Research Report No. 2007–2). New York, NY: College Entrance Examination Board.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339–355.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31–50. https://doi.org/10.1177/0265532209360671

Appendix A

```
title="4 subsets of 5 raters and 125 test-takers each"
facets=2           ; test-takers, raters
model=?,?,r4  ;  rating  scale  model  with  a  four-category
scale
labels=
1,Test-takers
1-125=Subset 1
126-250=Subset 2
251-375=Subset 3
376-500=Subset 4
*
2,Raters, G ; group-anchoring with group mean = 0
1-5=Subset 1,0,1
6-10=Subset 2,0,2
11-15=Subset 3,0,3
16-20=Subset 4,0,4
*
dvalue=2, 1-20 ; each test-taker can be rated by up to 20
raters
data=
;
1,  3,2,2,4,2,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.;  example  data
(test-taker  1  is  rated  by  raters  1-5,  missing  data  for
raters 6-20)
```