Educational and Psychological
Measurement
2018, Vol. 78(4) 679–707

DOI: 10.1177/0013164417703733
journals.sagepub.com/home/epm
SAGE

The Stabilizing Influences of Linking Set Size and Model–Data Fit in Sparse Rater-Mediated Assessment Networks

Stefanie A. Wind[1] and Eli Jones[2]

Abstract

Previous research includes frequent admonitions regarding the importance of establishing connectivity in data collection designs prior to the application of Rasch models. However, details regarding the influence of characteristics of the linking sets used to establish connections among facets, such as locations on the latent variable, model–data fit, and sample size, have not been thoroughly explored. These considerations are particularly important in assessment systems that involve large proportions of missing data (i.e., sparse designs) and are associated with high-stakes decisions, such as teacher evaluations based on teaching observations. The purpose of this study is to explore the influence of characteristics of linking sets in sparsely connected rating designs on examinee, rater, and task estimates. A simulation design whose characteristics were intended to reflect practical large-scale assessment networks with sparse connections were used to consider the influence of locations on the latent variable, model–data fit, and sample size within linking sets on the stability and model–data fit of estimates. Results suggested that parameter estimates for examinee and task facets are quite robust to modifications in the size, model–data fit, and latent-variable location of the link. Parameter estimates for the rater, while still quite robust, are more sensitive to reductions in link size. The implications are discussed as they relate to research, theory, and practice.

[1] The University of Alabama, Tuscaloosa, AL, USA
[2] The University of Missouri, Columbia, MO, USA

Corresponding Author:
Stefanie A. Wind, Educational Research, The University of Alabama, 313C Carmichael Hall, Tuscaloosa, AL 35487, USA.
Email: swind@ua.edu

Keywords

rater-mediated assessments, Rasch measurement, sparse designs, teacher evaluation

Performance assessments that involve rater judgments (i.e., rater-mediated assessments) are used across a variety of disciplines and settings, including in education to evaluate both student and teacher performance. A key feature of such assessments is their reliance on raters to provide estimates of the examinees' ability as measured by the performance task. This particular feature is of note because of the impact that individual rater characteristics may have on the scores that an examinee receives. The act of rating is a complex process that is inherently error-prone (Cronbach, 1990). Guilford (1954) noted that

> the use of ratings rests on the assumption that the human observer is a good instrument of quantitative observation, that he is capable of some degree of precision and some degree of objectivity. . . . While forced to have much confidence in quantitative human judgments, we must be ever alert to the weaknesses involved and to the many sources of personal biases in those judgments. (p. 278)

In light of these concerns, scoring procedures often include multiple raters for each performance to mediate the effect of differences in rater interpretations for individual examinees. In particular, methods based on Rasch measurement theory (Rasch, 1960) are often employed in performance assessments because they can be used to obtain estimates of examinee achievement that are adjusted for differences in rater severity (Eckes, 2015; Engelhard, 1994; Wind & Peterson, 2018). Rasch models are useful in the context of performance assessments because they do not require each rater to score each examinee performance so long as there are overlapping components in the assessment network, such as raters scoring common performances with other raters (i.e., connectivity; Eckes, 2015; Engelhard, 1997; Schumacker, 1999). Although approaches besides Rasch measurement models can also be used to adjust examinee estimates for differences in rater severity, such as the generalizability theory (e.g., Longford, 1994), as well as approaches based on analysis of variance (e.g., Braun, 1988) and regression (Lance, LaPointe, & Stewart, 1994; Raymond & Viswesvaran, 1993; Raymond, Webb, & Houston, 1991; Wilson, 1988), the current study focuses on issues associated with this procedure in the context of Rasch measurement theory.

Researchers have emphasized the need for connectivity in data collection designs to arrive at interpretable estimates when Rasch models are applied (Eckes, 2015; Engelhard, 1997; Myford & Wolfe, 2000; Schumacker, 1999; Wind, Engelhard, & Wesolowski, 2016). However, research related to the influence of the composition of links within incomplete assessment networks remains relatively inconclusive. In particular, characteristics of these linking sets, including sample size, judged proficiency, and the quality of ratings, have received limited attention in empirical analyses. Connections among facets in operational assessment systems are often sparse, meaning that connectivity is shared between a limited number of facets (Myford & Wolfe, 2000; Sykes, Ito, & Wang, 2008), which may lead to reduced precision in parameter estimates. In extreme cases of sparseness, large numbers of examinees may be rated by only a single rater, with raters sharing only a handful of ratings between them. Accordingly, it is essential to understand the influence of these characteristics, particularly in high-stakes assessment systems.

Sparse rating designs are prevalent across performance assessment contexts, including assessments of students (e.g., essay-based writing assessments), and teacher evaluation systems based on principal observations. In particular, teacher evaluations by principals reflect a growing area of concern in educational research and policy as these assessments are receiving increased weight in teacher evaluation procedures while lacking empirical support for their psychometric quality (e.g., Cohen & Goldhaber, 2016). Within the context of rater-mediated teacher evaluation systems based on principal observations, the current investigation seeks to provide additional insight into the effects of various characteristics of linking sets used to establish connectivity in sparse rating designs.

Purpose

The purpose of this study was to explore the influence of characteristics of linking sets in sparsely connected rating designs on examinee, rater, and task estimates within the context of rater-mediated teacher evaluation systems. Specifically, this study explored the influence of three characteristics of a linking set in terms of the degree to which manipulations of these characteristics resulted in changes in estimates of examinees, raters, and tasks: (1) the size of the link, (2) judged proficiency levels within the link, and (3) model–data fit within the link. Simulated data that reflect the characteristics of a large-scale teacher evaluation system based on principal observations were used to consider the following research questions:

1. What effect does the *size* of a linking set of complete ratings have on rater, examinee, and task parameter estimates in terms of location, model–data fit, and stability in sparse assessment networks?
2. What effect does the *model–data fit* within a linking set of complete ratings have on rater, examinee, and task parameter estimates in terms of location, model–data fit, and stability in sparse assessment networks?
3. What effect does the *judged proficiency level* within a linking set of complete ratings have on rater, examinee, and task parameter estimates in terms of location, model–data fit, and stability in sparse assessment networks?

In this study, we use the term *sparse assessment networks* to refer to rater-mediated assessment networks that are disconnected except for a relatively small linking set of examinees whose responses are scored by all of the raters.

Implications of Data Collection Designs in Rater-Mediated Assessments

Due to practical constraints, scoring procedures in operational rater-mediated assessments often involve incomplete designs. Accordingly, many researchers have explored the consequences of incomplete data collection designs on a variety of aspects of rater-mediated assessments. For example, within the framework of the generalizability theory, several scholars have described procedures for estimating rater reliability based on small subsets of complete ratings within incomplete assessment networks (Brennan, 2001; Chiu & Wolfe, 2002; DeMars, 2015). Essentially, these procedures are intended to provide estimates of rater reliability within the complete data set based on estimates of variance components within smaller subsets of complete ratings.

As noted above, Rasch models can also be applied in the context of incomplete ratings, as long as the data collection design includes connectivity among facets. A *connected* design is one in which every element is either directly or indirectly linked to all other elements (Eckes, 2015; Engelhard, 1997). When Rasch models are applied to connected designs, all of the facets (i.e., variables) in a rating system can be calibrated on a common scale (Eckes, 2015; Linacre & Wright, 2002); this result allows practitioners to make judgments about examinees' and raters' performance in relation to other examinees or raters. Although it is possible to obtain estimates based on the Rasch model in the presence of incomplete rating designs, it is important to note that large proportions of missing data within a rating design will result in large standard errors, which imply that the precision of these estimates may be reduced (Eckes, 2015).

Engelhard (1997) described a variety of rating designs based on incomplete data that result in data suitable for analysis with Rasch models. The key feature of these incomplete assessment networks is the presence of links, or common components between each facet in the assessment system. For example, in assessment systems that include raters and examinees, raters can score common examinee performances to provide direct and indirect links to other raters in the assessment network. Furthermore, when assessments include more than two facets, such as raters, examinees, and tasks, links can be established using common examinee performances on common tasks. In contrast, when systematic links are not included in rating designs, estimates of rater severity, examinee achievement, task difficulty, and other facet calibrations cannot be compared on a single linear continuum.

Myford and Wolfe (2000) explored the effects of various sample sizes, achievement levels, and rater consistency within linking sets of ratings in a large-scale rater-mediated speaking assessment and found that the stability of student achievement and rater severity estimates appeared to be related to the distribution of scores and rater consistency within linking sets. More recently, Wind et al. (2016) explored the consequences of differences in model–data fit for raters on student achievement estimates across rating designs with various levels of connectivity and found that the interpretation of these estimates depends on the degree to which raters demonstrate acceptable fit to the Rasch model. Although results from these studies suggest that the characteristics of common ratings used as links influence estimates based on incomplete ratings, the effects of these characteristics have not been explored systematically using a simulation study. Furthermore, the influence of linking set characteristics has not been explored previously in the context of rater-mediated teacher evaluation systems.

Method

This study uses simulated data to explore the influence of various characteristics of linking sets within sparse rating designs that reflect large-scale teacher evaluation systems based on principal observations. First, polytomous ratings were generated that were deliberately modified to reflect a sparse rating design, with only one rater scoring each examinee. The simulated data sets also contain a relatively small link of shared ratings common to all raters. Using a Monte Carlo simulation design, linking set characteristics are manipulated to explore the degree to which these characteristics correspond to changes in estimates of examinee achievement, rater severity, and task difficulty based on the Many-Facet Rasch (MFR) model.

To align the language of teacher evaluation systems with the language typically used in MFR model research—the following terms will be used to describe the variables in this study. Specifically, the *examinee* facet is made up of the teachers whose teaching performance is evaluated, the *rater* facet is made up of the principals who observe and evaluate teachers, and the *task* facet is made up of the aspects of teaching that are evaluated.

Simulation Design

A simulation study was conducted to systematically explore the consequences associated with manipulating linking set characteristics in sparsely connected rater-mediated assessment networks. In large-scale rater-mediated performance assessment systems based on sparse designs, such as teacher evaluation systems that involve principal observations, complete data (i.e., all raters score all examinees on all assessment components) do not exist. Accordingly, to provide insight into the influence of linking set characteristics in a manner that could be used to inform practice, the simulation was designed to match the characteristics of operational teacher evaluation systems based on principal observations.

First, it was necessary to generate complete sets of ratings from which ratings could be removed to create sparse designs that reflect practical constraints in operational large-scale rater-mediated assessment networks for teacher evaluation. Consequently, although the data generation procedure produced complete data, these fully crossed data sets were essentially an artifact of the simulation procedure that does not reflect practice. To reflect practical constraints in large-scale rater-mediated assessment procedures, ratings were removed from the complete data sets such that

Table 1. Simulation Design

	Design factors	Levels
Operational ratings	Rater sample size	50, 250
	Size of link[a]	3, 6, 8
Linking sets	Model–data fit within link[b]	Acceptable fit, noisy fit
	Average examinee measures within link[c]	Low, average, high

[a] Ratings were generated based on a linking set with n = 8 and modified to reflect the smaller linking set sample sizes (n = 6 and n = 3). [b] Acceptable fit was defined based on mean square error and standardized Rasch Outfit statistics following Smith, Schumacker, and Busch (1998): $1-(6/\sqrt{N}) \leq \text{Outfit}_{MSE} \leq 1+(6/\sqrt{N})$; $-2 \leq \text{Outfit}_{Z} < +2$. [c] Average examinee measures were characterized as follows: $M_\theta < -2$ = Low; $-1 < M_\theta < +1$ = Average; $M_\theta > +2$ = High.

the raters were fully disconnected with the exception of the linking set prior to any analyses. The resulting sparse data sets more closely reflect operational assessments in which fully crossed ratings do not exist.

Following the design specifications in Table 1, polytomous ratings were simulated based on a rating scale with five categories (0 = *low*; 4 = *high*) and the rating scale model (Andrich, 1978). The generating rater severity parameters were randomly selected from a uniform distribution between −4 and +4 logits, and generating examinee location parameters were simulated based on a standard normal distribution [$\theta \sim N(0,1)$]. The simulated data included four tasks with generating difficulty parameters of 0.00, −0.50, 0.50, and 1.00 logits, respectively. After the complete ratings were simulated, missingness was introduced by removing observations, such that each data set only included the selected number of examinees for each rater, and eliminating all common examinees between raters except for the linking sets. As a result, each simulated data set included two subsets: (1) fully disconnected operational ratings, in which each examinee was rated by only one rater with no connections among raters and (2) fully crossed linking set ratings.

Simulating Operational Ratings

First, operational ratings were generated that reflected two rater sample sizes (N = 50 or N = 250). Using the simulated data sets, observations were systematically removed to reflect fully disconnected rating designs. This systematic removal of observations was used to construct data sets that reflect operational settings in which complete data do not exist. First, the number of examinees that each rater scored was determined by randomly selecting a value between 4 and 40 from a uniform distribution. As a result, the number of examinees that each rater scored varied across raters within replications, as well as across replications. The range of possible values of examinees assigned to each rater was selected based on typical conditions observed within teacher evaluation systems, in which teacher observations are conducted across numerous schools with varying numbers of teachers and principals. Based on this specification, the range of examinee sample sizes for the conditions based on 50 raters could range from 200, if each rater scored 4 examinees, to 2,000, if each rater scored 40 examinees. Similarly, the range of examinee sample sizes for the conditions based on 250 raters could range from 1,000, if each rater scored 4 examinees, to 10,000 if each rater scored 40 examinees. Across conditions, only one rater scored each of the examinees in the operational rating sets. For each examinee, the rater provided scores on all four tasks.

Simulating Linking Set Ratings

Next, ratings were simulated to reflect linking sets that varied in terms of the size of the link (i.e., number of linked examinees; 3, 6, or 8), model–data fit for examinees within the link (acceptable fit or noisy fit), and logit-scale measures of the examinees within the links (low measures [$M_\theta = -2$]; average measures [$M_\theta = 0$], or high measures [$M_\theta = 2$]). All the raters scored all the examinees in the linking set, such that the ratings within the linking set were fully crossed.

The linking set sample sizes were selected based on the range of linking set sample sizes that are commonly observed in assessments based on observations of performance, including teacher evaluation systems (J. M. Linacre, personal communication, September 22, 2016). The simulation procedure resulted in acceptable model–data fit for each of the facets in the linking set; as a result, systemtatic manipulation was only needed to create the experimental conditions in which model–data fit was not acceptable. Because the misfit occurred at the examinee level, the misfit included in the linking set can be conceptualized as person misfit. In the context of item respose theory, person fit is most frequently examined in the context of selected-response assessments, where unexpected correct and incorrect responses for individual persons (e.g., as a result of guessing or testwiseness) contribute to response patterns described as person misfit. Although researchers have explored person fit less frequently in performance assessments compared with selected-response assessments (Cui & Mousavi, 2015; Meijer, Egberink, Emons, & Sijtsma, 2008; Rupp, 2013), unexpected responses at the examinee level can also occur in these assessments that result in person misfit. In the context of rater-mediated performance assessments, unexpected responses may occur as a result of inconsistent achievement across tasks for a particular examinee, or characteristics of an examinee's performance that result in inconsistent rater intepretations. Essentially, person misfit within the context of a rater-mediated performance assessment reflects a pattern of unexpected ratings associated with a particular performance.

A three-step procedure was used to incorporate person misfit into the linking set; this procedure is illustrated in the appendix and summarized breifly here. First, ratings were simulated such that data fit the model, with sample sizes and acheivement levels determined by the experimental condition (see Table 1). The data were structured such that each row reflected a unique examinee and each column reflected a unique rater within each task. Second, person misfit was introduced by creating discrepancies in rater severity ordering within the linking set. Specifically, within each task, the raters (columns) were reordered such that the relative severity ordering of each of the raters within the linking set was different from the relative severity ordering observed within the operational ratings. The new sequence for the raters within the linking set was determined using random sampling without replacement. Finally, the column labels for the Rater ID numbers were replaced to reflect the original rater ordering, while retaining the contents of the rearranged columns. The rater reordering within the linking set resulted in a different relative rater severity ordering between the operational and linking set ratings when the operational and linking set ratings were analyzed together. These discrepancies resulted in higher-than-expected values of model–data fit statistics within the linking set. Because high values of the Rasch Outfit statistic, which indicate extreme residuals or unexpected responses, are generally described as more cause for concern than low values, which indicate less variation than expected by the probabilistic model (Wolfe & Smith, 2007), the simulation procedures for introducing misfit were limited to higher-than-expected values of Outfit statistics (i.e., ‘‘noisy’’ fit).

After they were generated, the linking set ratings were appended to the operational ratings to create sparse rating designs in the simulated data. One hundred replications were completed within each cell of the simulation design.

Data Analysis

The simulated data were analyzed using a two-step procedure. First, the rating scale formulation of the MFR model (Andrich, 1978; Linacre, 1989) was applied to each data set using the Facets computer program (Linacre, 2015). The rating scale formulation of the MFR model was selected to match previous research on data collection systems for rater-mediated assessments (Myford & Wolfe, 2000; Wind et al., 2016). Stated mathematically, the model is as follows:

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

where θ_n represents the judged achievement level of examinee n on the logit scale; λ_i represents the severity level of rater i on the logit scale; δ_j represents the judged difficulty of task j on the logit scale, and τ_k is the location on the logit scale where the probability for a rating in category k and a rating in category $k-1$ are equal.

After estimates for each facet were obtained for the simulated data sets, three major dependent variables were of interest: (1) logit-scale locations for elements within facets and their corresponding standard errors, (2) estimates of model–data fit for elements within facets, and (3) correlations between examinee location estimates based on the complete linking set conditions and the estimates based on the modified linking set conditions. These dependent variables are elaborated below.

Logit-Scale Locations and Standard Errors

First, logit-scale locations and their corresponding standard errors were examined for the examinee, rater, and task facets across each of the simulated data sets. In particular, the average location and spread of estimates (i.e., standard deviation) of these estimates were examined in order to consider differences in the overall calibrations and variation in facet locations across conditions. Standard errors for each facet were also examined as an indicator of the precision of the logit-scale estimates.

Model–Data Fit

Second, estimates of model–data fit were examined for the examinee, rater, and task facets across each of the simulated data sets. Specifically, values of the unstandardized (mean square error [*MSE*]) and standardized infit and outfit statistics were examined for each of the facets across each of the conditions to consider differences in the overall alignment between model expectations and observed response patterns related to differences in connectivity.

Correlations Across Linking Set Conditions

Finally, the stability of examinee estimates across data collection designs was examined using correlations between each of the original linking set conditions ($N_{\text{Link}} = 8$) and the corresponding conditions based on modified linking set sample sizes. Specifically, these correlations were calculated between examinee estimates from each replication of the simulation conditions based on linking sets with eight common examinees and the corresponding estimates based on modifications of these data sets that included six or three common examinees in the linking sets. This correlation procedure reflects the use of the simulation conditions based on linking sets with eight common examinees ($N_{\text{Link}} = 8$) as the frame of reference for considering the influence of different acheivement levels and model–data fit across linking sets with smaller sample sizes. Accordingly, it should be noted that the correspondence between the original linking set conditions ($N_{\text{Link}} = 8$) and the two modified sample sizes ($N_{\text{Link}} = 6$ and $N_{\text{Link}} = 3$) was more meaningful than the correspondence between the observed estimates and the original generating parameters for the complete ratings. Because we were interested in the changes in examinee, rater, and task estimates across sample sizes in the presence of different achievement levels and model–data fit, the degree to which the estimation procedure recovered the original parameters for the complete ratings (i.e., bias in the initial parameter estimates for $N_{\text{Link}} = 8$) was beyond the scope of the study. This point is discussed further at the end of the article.

Results

In this section, the results are organized as follows. First, descriptive statistics for the conditions based on the original linking set sample size ($N_{\text{Link}} = 8$) are presented to verify that the simulation procedure resulted in ratings with the intended characteristics and to establish the characteristics of the conditions that serve as a frame of reference for the modified linking set sample sizes. Then, results are summarized according to the three linking set characteristics that correspond to the guiding research questions for this study: (1) size of the linking set, (2) model–data fit within the linking set, and (3) judged proficiency level within the linking set. Conclusions and a discussion of the results in terms of the research questions follow.

Calibration of the Linking Sets

To verify that the simulated ratings matched the intended design specifications, summary statistics were examined within the linking sets and operational ratings across replications for all of the simulation conditions based on the original linking set sample size (N_{Link} = 8). Within the linking sets, the characteristics of the examinee facet were of particular interest because this facet was manipulated in the simulation conditions. Table 2 includes summary statistics for the examinee facet within the linking sets based on eight common examinees. The results in Table 2 support the hypothesis that the simulated ratings reflect the intended characteristics. Specifically, values of the location estimates were lowest in the conditions in which the generating parameters were specified as low, followed by the conditions in which the generating parameters were specified as average, and highest in the conditions in which the generating parameters were specified as high. The standard errors associated with the logit-scale locations suggest that the estimates of examinee locations within the linking set were more precise (i.e., smaller standard errors [*SE*]) in the conditions based on the larger rater sample size (N_{Raters} = 250: $0.04 \leq SE \leq 0.06$) compared with the conditions based on the smaller rater sample size (N_{Raters} = 50: $0.10 \leq SE \leq 0.14$). Furthermore, the average standard errors suggested slightly less precise estimates for linking set examinees within the conditions based on acceptable model–data fit compared with the conditions based on noisy model–data fit. Finally, values of the model–data fit statistics were higher and exceeded the critical values in the conditions in which the linking set was specified as noisy than in the conditions in which the linking set was specified to have acceptable model–data fit.

Calibration of the Operational Ratings

Next, summary statistics were examined within the operational ratings for the conditions in which the original linking set sample size was used (N_{Link} = 8). These results are summarized in Tables 3, 4, and 5 for the examinee, rater, and task facets, respectively. Across the three facets, the results suggested that the overall calibration of examinees, raters, and tasks matched the specifications of the simulation design, with average estimated logit-scale locations around zero for each facet.

In terms of standard errors, the results suggested notable differences related to the precision of logit-scale locations between the operational ratings and the linking set

Table 2. Summary Statistics Within Linking Sets Across Simulation Conditions: Examinee Facet: $N_{\text{Link}} = 8$

Rater sample size	Specified location within linking set	Specified model–data fit within linking set	Measure		Standard error		Model–data fit							
							Infit MSE		Std. infit		Outfit MSE		Std. outfit	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Low	Acceptable	−2.78	1.21	0.14	0.03	1.12	0.15	0.79	0.95	0.97	0.33	−0.04	0.68
		Noisy	−2.71	1.09	0.12	0.03	1.49	0.19	3.16	1.23	1.28	0.31	1.07	0.94
	Average	Acceptable	−0.34	1.23	0.11	0.01	1.11	0.13	0.88	1.02	1.09	0.24	0.49	0.96
		Noisy	−0.31	1.05	0.10	0.01	1.40	0.14	3.11	1.03	1.39	0.21	2.32	1.14
	High	Acceptable	2.18	1.21	0.12	0.02	1.11	0.13	0.78	0.94	0.99	0.23	0.01	0.71
		Noisy	2.18	1.03	0.11	0.02	1.46	0.18	3.16	1.16	1.30	0.27	1.31	1.00
250	Low	Acceptable	−2.73	1.16	0.06	0.01	1.12	0.07	1.90	1.01	0.99	0.18	−0.14	0.75
		Noisy	−2.72	1.04	0.06	0.01	1.51	0.11	7.18	1.77	1.30	0.15	2.31	1.58
	Average	Acceptable	−0.34	1.14	0.05	0.00	1.13	0.06	2.30	1.06	1.12	0.09	1.40	1.01
		Noisy	−0.26	0.98	0.04	0.00	1.41	0.07	7.03	1.06	1.39	0.09	5.18	1.43
	High	Acceptable	2.28	1.13	0.06	0.01	1.11	0.06	1.81	0.99	1.00	0.15	−0.09	0.82
		Noisy	2.18	0.96	0.05	0.01	1.48	0.09	7.27	1.40	1.31	0.12	2.90	1.58

Note. MSE = mean square error; Std. = standardized; *M* = mean; *SD* = standard deviation.

Table 3. Summary Statistics Across Simulation Conditions: Examinee Facet (Operational Only); $N_{\text{Link}} = 8$

							Model–data fit							
Rater sample size	Specified location within linking set	Specified model–data fit within linking set	Measure		Standard error		Infit MSE		Std. infit		Outfit MSE		Std. outfit	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Low	Acceptable	−0.36	1.54	0.98	0.47	0.90	0.62	−0.06	0.88	0.93	0.70	−0.02	0.87
		Noisy	−0.31	1.34	0.94	0.49	0.82	0.51	−0.15	0.82	0.81	0.50	−0.15	0.80
	Average	Acceptable	−0.28	1.47	0.97	0.47	0.89	0.62	−0.08	0.88	0.92	0.69	−0.04	0.87
		Noisy	−0.26	1.27	0.92	0.48	0.81	0.50	−0.16	0.83	0.80	0.49	−0.16	0.81
	High	Acceptable	−0.26	1.52	0.98	0.48	0.89	0.62	−0.07	0.87	0.92	0.69	−0.03	0.87
		Noisy	−0.20	1.32	0.93	0.48	0.82	0.50	−0.16	0.82	0.80	0.49	−0.16	0.80
250	Low	Acceptable	−0.32	1.50	0.98	0.47	0.90	0.62	−0.06	0.88	0.93	0.70	−0.02	0.87
		Noisy	−0.29	1.31	0.93	0.48	0.82	0.51	−0.15	0.83	0.81	0.51	−0.15	0.81
	Average	Acceptable	−0.26	1.43	0.97	0.47	0.89	0.61	−0.08	0.87	0.91	0.69	−0.04	0.87
		Noisy	−0.23	1.24	0.93	0.49	0.81	0.49	−0.16	0.82	0.80	0.48	−0.16	0.80
	High	Acceptable	−0.23	1.49	0.97	0.47	0.90	0.62	−0.07	0.87	0.92	0.69	−0.03	0.87
		Noisy	−0.22	1.29	0.93	0.49	0.81	0.50	−0.16	0.82	0.80	0.49	−0.16	0.80

Note. *MSE* = mean square error; Std. = standardized; *M* = mean; *SD* = standard deviation. The average examinee sample size for the $N_{\text{Raters}} = 50$ conditions was 1093.15 (*SD* = 9.24); the average examinee sample size for the $N_{\text{Raters}} = 250$ conditions was 5499.90 (*SD* = 12.63).

Table 4. Summary Statistics Across Simulation Conditions: Rater Facet; $N_{\text{Link}} = 8$

Rater sample size	Specified location within linking set	Specified model–data fit within linking set	Measure		Standard error		Model–data fit							
							Infit MSE		Std. infit		Outfit MSE		Std. outfit	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Low	Acceptable	0.00	2.15	0.18	0.09	0.95	0.19	−0.23	1.00	0.94	0.40	−0.02	0.83
		Noisy	0.00	1.93	0.16	0.08	0.97	0.24	−0.15	1.30	0.97	0.36	0.09	1.03
	Average	Acceptable	0.00	2.35	0.16	0.06	0.96	0.17	−0.21	0.98	0.98	0.25	−0.10	0.89
		Noisy	0.00	1.95	0.15	0.05	0.99	0.18	−0.06	1.01	1.01	0.24	0.00	0.96
	High	Acceptable	0.00	2.16	0.16	0.06	0.96	0.18	−0.24	1.01	0.94	0.27	−0.10	0.84
		Noisy	0.00	1.91	0.15	0.06	0.99	0.25	−0.10	1.32	0.98	0.30	0.04	1.02
250	Low	Acceptable	0.00	2.18	0.17	0.08	0.96	0.19	−0.20	1.03	0.95	0.37	−0.01	0.86
		Noisy	0.00	1.95	0.16	0.08	0.98	0.25	−0.14	1.33	0.98	0.36	0.11	1.04
	Average	Acceptable	0.00	2.36	0.16	0.06	0.97	0.17	−0.18	0.98	0.98	0.26	−0.06	0.90
		Noisy	0.00	2.00	0.15	0.05	0.99	0.19	−0.07	1.04	1.01	0.22	0.00	0.97
	High	Acceptable	0.00	2.17	0.16	0.06	0.96	0.18	−0.22	1.01	0.95	0.32	−0.08	0.86
		Noisy	0.00	1.92	0.15	0.06	1.00	0.25	−0.09	1.37	0.99	0.31	0.05	1.04

Note. *MSE* = mean square error; Std. = standardized; *M* = mean; *SD* = standard deviation.

Table 5. Summary Statistics Across Simulation Conditions: Task Facet; $N_{\text{Link}} = 8$

							Model–data fit							
Rater sample size	Specified location within linking set	Specified model–data fit within linking set	Measure		Standard error		Infit MSE		Std. infit		Outfit MSE		Std. outfit	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Low	Acceptable	0.00	0.71	0.04	0.00	0.97	0.04	−0.70	0.87	0.94	0.08	−0.62	0.81
		Noisy	0.00	0.36	0.04	0.00	0.98	0.13	−0.48	2.89	0.96	0.21	−0.87	2.25
	Average	Acceptable	0.00	0.69	0.04	0.00	0.97	0.04	−0.73	0.86	0.97	0.06	−0.48	0.94
		Noisy	0.00	0.27	0.04	0.00	0.99	0.14	−0.41	3.19	0.98	0.15	−0.47	2.78
	High	Acceptable	0.00	0.69	0.04	0.00	0.96	0.04	−0.83	0.85	0.94	0.06	−0.75	0.77
		Noisy	0.00	0.32	0.04	0.00	0.99	0.14	−0.47	3.10	0.96	0.20	−0.87	2.57
250	Low	Acceptable	0.00	0.70	0.02	0.00	0.97	0.02	−1.52	1.02	0.95	0.04	−1.33	1.12
		Noisy	0.00	0.36	0.02	0.00	0.99	0.13	−1.11	6.30	0.96	0.22	−1.83	4.82
	Average	Acceptable	0.00	0.68	0.02	0.00	0.97	0.02	−1.54	1.12	0.97	0.03	−1.01	1.14
		Noisy	0.00	0.27	0.02	0.00	0.99	0.13	−0.92	6.94	0.98	0.14	−1.15	5.99
	High	Acceptable	0.00	0.69	0.02	0.00	0.96	0.02	−1.79	1.01	0.95	0.03	−1.54	1.05
		Noisy	0.00	0.32	0.02	0.00	0.99	0.13	−1.10	6.68	0.96	0.19	−2.06	5.54

Note. MSE = mean square error; Std. = standardized; *M* = mean; *SD* = standard deviation.

ratings, as well as across the three facets. Specifically, the average standard errors for the examinee facet were substantially larger within the operational ratings compared with those observed within the linking set (see Table 2)—suggesting less precision in operational examinee estimates based on the sparse rating design. However, the standard errors for the examinee facet were comparable across the two rater sample sizes, with slightly larger standard errors (i.e., slightly less precision) observed within the conditions based on acceptable model–data fit within the linking set compared with the conditions based on noisy model–data fit. Compared with the examinee facet, the standard errors were smaller for the rater and task facets—indicating more precise logit-scale location estimates for individual raters and tasks. The standard errors for the rater facets also showed the pattern that the estimates obtained from the conditions with acceptable model–data fit within the linking set were more precise than those obtained from the conditions based on noisy model–data fit. However, for the task facet, average values of the standard errors were equivalent across conditions.

Finally, average values of all four model–data fit statistics reflect adequate overall fit for the examinee, rater, and task facets. However, within the rater and task facets, slightly more variation was observed among values of the standardized Infit and Outfit statistics within the conditions based on linking sets with noisy fit.

Size of the Linking Set

The first research question focuses on the influence of the sample size within the linking set on parameter estimates, model–data fit, and the stability of estimate in sparse rating designs. Because the conditions based on a linking set with eight common examinees served as the frame of reference, the results in this section focus on the conditions in which the size of the linking set was modified from eight common examinees to six or three common examinees.

Parameter Estimates. First, logit-scale locations and corresponding standard errors were examined for each facet across the conditions based on the two modified linking set sample sizes. For examinees, the average estimated proficiency levels were fairly stable across conditions with linking sets based on six and three common examinees ($-0.36 \leq M_\theta \leq -0.22$). Although these estimates reveal a slight downward bias between the observed average examinee locations and the generating parameters, which were specified with a mean of zero logits, the finding of generally consistent average estimates between the average estimates within the conditions based on the original linking set sample size (see Table 3) and the two modified sample sizes suggests that the examinee location estimates were generally robust across modifications to the linking sets. Furthermore, the average standard errors for the examinee facet based on the two modified linking sample sizes were comparable with the average values observed within the conditions based on the original linking set sample size (see Table 3). Specifically, for both modified linking set sample size conditions, the average standard errors for examinees were comparable across the two rater sample sizes ($N_{\text{Raters}} = 50$: $0.92 \leq M_{SE,\theta} \leq 0.99$; $N_{\text{Raters}} = 250$: $0.93 \leq M_{SE,\theta} \leq 0.98$).

For raters, the average severity calibrations were stable across conditions with linking sets based on six and three common examinees ($M_{\lambda} = 0.00$ across conditions). The average standard errors for the rater facet based on the two modified linking sample sizes were also comparable to the average values observed within the conditions based on the original linking set sample size (see Table 4) and across the two rater sample sizes ($N_{\text{Raters}} = 50$: $0.16 \leq M_{SE,\lambda} \leq 0.18$; $N_{\text{Raters}} = 250$: $0.16 \leq M_{SE,\lambda} \leq 0.18$).

Similar to raters, the average task difficulty calibrations were comparable across conditions with linking sets based on six and three common examinees ($M_{\delta} = 0.00$ across conditions). The average standard errors for the task facet based on the two modified linking sample sizes were also comparable to the values observed based on the original linking set sample size. Across both modified linking set sample sizes, the average values of the standard error for the task estimates were comparable across the two rater sample sizes ($N_{\text{Raters}} = 50$: $0.04 \leq M_{SE,\delta} \leq 0.05$; $N_{\text{Raters}} = 250$: $M_{SE,\delta} = 0.02$).

Model–Data Fit. Next, values of Rasch model–data fit statistics were examined for the examinee, rater, and task facets across the conditions in which the size of the linking set was modified from eight common examinees to six or three common examinees. For all three facets, acceptable overall average values of unstandardized and standardized fit statistics were observed based on all the conditions in which the linking sets were modified to include either six or three common examinees. These values were quite stable across the modified linking set sizes for examinees ($\Delta_{\text{Infit } MSE} \leq 0.08$; $\Delta_{\text{Std. Infit}} \leq 0.08$; $\Delta_{\text{Outfit } MSE} \leq 0.12$; $\Delta_{\text{Std. Outfit}} \leq 0.11$) and raters ($\Delta_{\text{Infit } MSE} \leq 0.02$; $\Delta_{\text{Std. Infit}} \leq 0.06$; $\Delta_{\text{Outfit } MSE} \leq 0.03$; $\Delta_{\text{Std. Outfit}} \leq 0.09$). However, slightly more variation was observed among the model–data fit statistics for the task facet between linking set sizes, particularly with regard to the standardized fit statistics ($\Delta_{\text{Infit } MSE} \leq 0.01$; $\Delta_{\text{Std. Infit}} \leq 0.35$; $\Delta_{\text{Outfit } MSE} \leq 0.02$; $\Delta_{\text{Std. Outfit}} \leq 0.65$), where noisier fit statistics (higher values of fit statistics) were observed within the conditions based on $N_{\text{Link}} = 3$.

Stability of Estimates. Next, bivariate correlations were calculated between the logit-scale locations of examinees, raters, and tasks based on the conditions that included eight common examinees and the estimates based on the modified sample sizes ($N_{\text{Link}} = 6$ and $N_{\text{Link}} = 3$). Specifically, within each condition, estimates of examinee proficiency within the operational ratings based on $N_{\text{Link}} = 6$ or $N_{\text{Link}} = 3$ were correlated with estimates of examinee proficiency obtained from corresponding conditions based on $N_{\text{Link}} = 8$. Average values of the correlation coefficients and standard deviations across the replications of each condition are presented in Table 6.

For the examinee facet, results from the correlation analysis suggest very high correspondence ($r \geq 0.96$) between estimates based on the original linking set

Table 6. Average Correlations Between Estimates Based on Original and Modified Linking Set Sample Sizes (Average Across Replications).

Rater sample size	Specified location within linking set	Specified model–data fit within linking set	Examinees				Raters				Tasks			
			$r_{(NLink = 8, NLink = 6)}$		$r_{(NLink = 8, NLink = 3)}$		$r_{(NLink = 8, NLink = 6)}$		$r_{(NLink = 8, NLink = 3)}$		$r_{(NLink = 8, NLink = 6)}$		$r_{(NLink = 8, NLink = 3)}$	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	Low	Acceptable	0.99	0.01	0.96	0.03	0.94	0.23	0.93	0.23	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	0.99	0.00	0.96	0.02	0.82	0.06	1.00	0.00	1.00	0.00
	Average	Acceptable	0.99	0.00	0.96	0.01	1.00	0.00	0.99	0.01	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	1.00	0.00	0.96	0.01	0.83	0.05	1.00	0.00	1.00	0.00
	High	Acceptable	0.99	0.01	0.96	0.02	1.00	0.00	0.98	0.01	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	0.99	0.00	0.95	0.02	0.80	0.08	1.00	0.00	1.00	0.00
250	Low	Acceptable	0.99	0.01	0.96	0.02	0.97	0.13	0.96	0.14	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	0.99	0.00	0.89	0.25	0.80	0.23	1.00	0.00	1.00	0.00
	Average	Acceptable	0.99	0.00	0.97	0.01	0.98	0.11	0.97	0.12	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	1.00	0.00	0.98	0.00	0.90	0.01	1.00	0.00	1.00	0.00
	High	Acceptable	0.99	0.00	0.97	0.01	1.00	0.00	0.98	0.01	1.00	0.00	1.00	0.00
		Noisy	1.00	0.00	0.99	0.00	0.91	0.23	0.82	0.22	1.00	0.00	1.00	0.00

Note. M = mean; *SD* = standard deviation.

sample size (N_{Link} = 8) and the two modified sample sizes, with average correlations between the original sample size and the conditions based on six common examinees slightly higher and less variable than the correlations between the original sample size and the conditions based on three common examinees. For the rater facet, high correlations ($r \geq$ 0.89) were also observed between estimates based on the original linking set sample size and the conditions based on six common examinees for both rater sample sizes. Values of the correlation coefficient were slightly lower for the conditions based on three common examinees (0.80 $\leq r \leq$ 0.98). Correlations between task estimates based on the original linking set sample size and the conditions based on six and three common examinees reflect the general pattern observed for the examinee and rater facets. Specifically, very high correlations ($r \geq$ 0.99) were observed between estimates based on the original linking set sample size and the conditions based on six and three common examinees for both rater sample sizes.

Model–Data Fit Within the Linking Set

The second research question focuses on the influence of differences in model–data fit within the linking set on parameter estimates, model–data fit, and the stability of estimates.

Parameter Estimates. Although the overall calibration of the three facets was comparable across model–data fit conditions, some differences were observed with regard to the spread of logit-scale locations. For examinees, slightly higher values of standard deviations were observed within the conditions based on N_{Link} = 6 when the linking sets included acceptable model–data fit (1.45 $\leq SD_\theta \leq$ 1.58) compared with the conditions based on linking sets with noisy model–data fit (N_{Link} = 6: 1.28 $\leq SD_\theta \leq$ 1.39). However, for the conditions based on N_{Link} = 3, the examinee estimates were more variable, particularly in the conditions based on 250 raters. Whereas the standard deviations of examinee estimates when N_{Raters} = 50 reflected the general pattern of higher standard deviations for the conditions based on acceptable model–data fit (1.55 $\leq SD_\theta \leq$ 1.69) than the conditions based on noisy model–data fit (1.41 $\leq SD_\theta \leq$ 1.52), the opposite pattern was observed when N_{Raters} = 250: (acceptable model–data fit: 1.50 $\leq SD_\theta \leq$ 1.65; noisy model–data fit: 2.77 $\leq SD_\theta \leq$ 2.78). In terms of standard errors, slightly higher average standard errors, which suggest less precise estimates, were observed for the examinee estimates within conditions based on acceptable model–data fit (0.92 $\leq M_{SE,\theta} \leq$ 0.95) compared with the conditions based on noisy model–data fit (0.97 $\leq M_{SE,\theta} \leq$ 0.99).

For the rater facet, slightly larger standard deviations in rater severity calibrations were observed based on linking sets with acceptable model–data fit (2.10 $\leq SD_\lambda \leq$ 2.37) compared with the conditions in which the linking sets included noisy model–data fit (1.91 $\leq SD_\lambda \leq$ 2.05). Similar to examinees, slightly higher average standard errors for rater severity calibrations were observed for the conditions based on acceptable model–data fit (0.17 $\leq M_{SE,\lambda} \leq$ 0.18) compared with the conditions based on noisy model–data fit ($0.16 \leq M_{SE,\lambda} \leq 0.17$). For tasks, slightly larger standard deviations among task difficulty calibrations were observed based on linking sets with acceptable model–data fit (N_{Link} = 6: $0.69 \leq SD_\delta \leq 0.72$; N_{Link} = 3: $0.71 \leq SD_\delta \leq 0.74$) compared with the conditions in which the linking sets included noisy model–data fit (N_{Link} = 6: $0.34 \leq SD_\delta \leq 0.42$; N_{Link} = 3: $0.53 \leq SD_\delta \leq 0.70$). Furthermore, slightly higher average standard errors for the task estimates were observed for the conditions based on acceptable model–data fit ($0.02 \leq M_{SE,\delta} \leq 0.05$) compared with the conditions based on noisy model–data fit ($0.02 \leq M_{SE,\delta} \leq 0.04$).

Model–Data Fit. For examinees, acceptable average values across replications of unstandardized and standardized fit statistics were observed across conditions based on linking sets with acceptable and noisy model–data fit, where the maximum absolute difference in Infit *MSE* or Outfit *MSE* across conditions based on linking sets with acceptable and noisy model–data fit was 0.12 and the maximum change in standard deviations for Infit and Outfit *MSE* was 0.20. With regard to standardized fit statistics, the maximum absolute difference in average standardized Infit or Outfit across conditions based on linking sets with acceptable and noisy model–data fit was 0.13, and the maximum change in standard deviations for standardized Infit and Outfit was 0.07.

For the rater facet, model–data fit statistics were slightly more variable when the linking sets included noisy model–data fit based on both modified linking set conditions, particularly with regard to standardized fit. Specifically, the maximum absolute difference in rater Infit *MSE* or Outfit *MSE* across conditions based on linking sets with acceptable and noisy model–data fit was 0.05 and the maximum change in standard deviations for Infit and Outfit *MSE* was 0.09, where higher values of model–data fit statistics and higher standard deviations were observed within the conditions based on noisy model–data fit. In terms of standardized fit statistics, the maximum absolute difference in Infit or Outfit across conditions based on linking sets with acceptable and noisy model–data fit was 0.15, and the maximum change in standard deviations for standardized Infit and Outfit was 0.34; similar to the unstandardized rater fit statistics, higher values of model–data fit statistics and higher standard deviations were observed within the conditions based on noisy model–data fit.

For the task facet, average values of each of the fit statistics were generally within the range of expected values when data fit the Rasch model across both modified linking set conditions. However, several large standard deviations appeared for the standardized Infit and Outfit statistics within conditions based on linking sets with noisy model–data fit. Specifically, the maximum absolute difference in task Infit *MSE* or Outfit *MSE* across conditions based on linking sets with acceptable and noisy model–data fit was 0.02, and the maximum change in standard deviations for Infit and Outfit *MSE* was 0.10, where higher values of model–data fit statistics and higher standard deviations were observed within the conditions based on noisy model–data fit. The maximum absolute difference in standardized Infit or Outfit across conditions based on linking sets with acceptable and noisy model–data fit was 2.22, and the maximum change in standard deviations for standardized Infit and Outfit was 4.36, where higher values of model–data fit statistics and higher standard deviations were observed within the conditions based on noisy model–data fit.

Stability of the Estimates. For examinees, correlations based on conditions in which the linking set was noisy were slightly higher ($0.99 \leq M_r \leq 1.00$) than the correlations based on conditions in which the linking set included acceptable model–data fit ($0.96 \leq M_r \leq 0.99$). Although the overall average correlations were lower for the rater facet, a similar pattern was observed with regard to the model–data fit conditions. Specifically, average correlations between rater severity calibrations were higher in the conditions in which the linking sets displayed adequate model–data fit ($0.80 \leq M_r \leq 0.96$) compared with those with noisy fit ($0.93 \leq M_r \leq 1.00$); this pattern persisted across rater sample size conditions. For tasks, the correlations were equivalent across model–data fit conditions ($M_r = 1.00$, $SD = 0.00$).

Judged Proficiency Level Within the Linking Set

The third research question focuses on the influence of the judged achievement level within the linking set on parameter estimates, model–data fit, and the stability of estimate in sparse rating designs. The results from the simulation study revealed that the values of parameter estimates and corresponding standard errors, model–data fit statistics, and stability of the estimates were comparable across the simulation conditions in which the linking set was manipulated to reflect low, average, and high achievement levels.

Summary and Conclusions

The purpose of this study was to explore the influence of characteristics of linking sets in sparse rating designs on examinee, rater, and task estimates. Simulated data that reflect the characteristics of large-scale teacher evaluation systems based on principal observations were used to consider the effects of embedding linking sets with varying characteristics within sparsely connected assessment networks. Specifically, differences in sample size, judged proficiency level, and model–data fit within linking sets were considered in terms of their influence on estimates of examinee (teacher) proficiency, rater (principal) severity, and task difficulty. Overall, the results suggested that embedding linking sets of complete ratings within disconnected assessment networks facilitated the calibration of examinees, raters, and tasks on a common scale and that estimates of examinee proficiency, rater severity, and task difficulty were relatively stable across linking sets with different characteristics. In this section, tentative conclusions are presented as they relate to the research questions.

What effect does the size of a linking set of complete ratings have on estimates of examinee proficiency, rater severity, and task difficulty in terms of location, model–data fit, and stability in sparse assessment networks?

Parameter Estimates. In terms of parameter estimates for examinees, raters, and tasks, results from the simulation study suggested that parameter estimates were not substantially affected by changes in the size of the link. Changes in average parameter estimates across all conditions were marginal ($0.01 \leq \Delta M \leq 0.06$). Furthermore, the overall spread of rater and task calibrations on the logit scale was comparable across linking set sizes.

With regard to standard errors for the parameter estimates, we observed a large difference in the precision of the logit-scale location estimates for examinees between the linking set and operational ratings, where the estimates of examinees within the linking set were more precise (smaller standard errors) compared with those in the operational set. However, this result was to be expected based on previous research on sparse rating designs. Furthermore, the values of standard errors were comparable across modifications of the linking set sample size.

Model–Data Fit. Results suggested that model–data fit for examinees remained acceptable across varying sizes of the linking set. However, slightly more variation was observed in model–data fit for the task facet when the sample size within the linking set was reduced. These results suggest that model–data fit may be more variable when a smaller link size is used.

Stability of Estimates. Results suggested a high degree of stability in estimated examinee proficiency measures and task difficulty measures between the original linking set sample sizes and the two modified linking set sample sizes. Only slightly more variability was observed when the linking set included three common examinees.

Similarly, strong positive correlations for rater calibrations between the original linking set sample sizes and the two modified linking set sample sizes indicate that rater severity measures also had a high level of stability. However, these correlations were markedly lower than those of the other two facets and also exhibited more variability. This finding suggests that the relative ordering of raters changed slightly when the size of the linking set was reduced. Furthermore, the rater calibrations were slightly less consistent when the linking set included three common examinees than when the linking set included six common examinees.

What effect does model–data fit within a linking set of complete ratings have on rater, examinee, and task parameter estimates in terms of location, model–data fit, and stability in sparse assessment networks?

Parameter Estimates. The results suggest that the linking sets with additional noise tended to provide more consistent and precise estimates of facet parameters than linking sets with more acceptable fit levels, where smaller standard deviations in logit-scale locations and smaller standard errors were observed when the linking set included noisy model–data fit than when the linking set included acceptable fit. These results suggest that more noise in the ratings of the common examinees that are included in the link may result in more information about facets in the rating design.

Model–Data Fit. Overall model–data fit for the examinees and raters remained acceptable for both noisy and acceptable linking set conditions. However, slightly more variation was observed in model–data fit for the task facet when noisy linking sets were embedded in the disconnected assessment networks. The increase in variation suggests that reduction in sample size may provide slightly less consistent estimates of task difficulty.

Stability of Estimates. The correlations between parameter estimates obtained from both acceptable and noisy linking sets suggested slightly more stable parameter estimates for examinees when the linking set included noisy model–data fit. For the rater facet, conditions based on noisy fit within the linking set resulted in slightly less stable parameter estimates. For the task facet, the results did not suggest any noticeable impact on the stability of parameter estimates for the task facet.

What effect does the judged proficiency level *within a linking set of complete ratings have on rater, examinee, and task parameter estimates in terms of location, model–data fit, and stability in sparse assessment networks?*

The results did not indicate that judged proficiency level within the link had a meaningful effect on any of the estimated facet parameters. This result was true for location and variability of parameter estimates, model–data fit, and stability, suggesting that judged proficiency level may not noticeably contribute to the quality of the link.

Discussion

The results from this study support the idea that disconnected rating designs can be successfully connected using a block of shared ratings (Eckes, 2015; Engelhard, 1997; Myford & Wolfe, 2000). Connecting a disconnected data set with a block of shared ratings can result in stable estimates of examinee ability and task difficulty and provide useful information regarding rater severity. This finding is important because sparsely connected rating designs are frequently used in operational assessment situations, including teacher evaluation systems based on principal observations. Practitioners and researchers can obtain quality estimates of examinee ability by using small subsets of fully crossed ratings within sparsely connected assessment systems. However, estimates for principal severity and other characteristics may be slightly less precise.

When considering the results from this study in terms of previous research, it is interesting to note that the sample size within the linking set appeared to have a more notable impact on the calibration and stability of the parameter estimates compared with the other manipulated characteristics. In particular, the finding that logit-scale locations within the linking set did not have a notable impact stands somewhat in contrast to the results reported by Myford and Wolfe (2000), who reported that higher logit-scale locations within linking sets appeared to result in higher quality connections within incomplete rating designs. On the other hand, our results were similar to those reported by Myford and Wolfe (2000) related to model–data fit within linking sets. Specifically, these authors reported that linking sets made up of less-consistent (i.e., more misfitting) benchmark performances resulted in higher levels of stability across rating designs than did linking sets made up of more-consistent performances.

Implications for Practice

The findings from this study highlight several important considerations for practitioners and researchers when designing data collection systems for rater-mediated performance assessments. First, these findings suggest that even small links may provide relatively stable measures of examinee ability and task difficulty and, to a lesser extent, rater severity. This finding has important implications for practice because using a shared block of ratings in situations where a fully, or even moderately, connected design is too expensive or impractical can provide a feasible way to facilitate the interpretation of examinee achievement measures and task difficulty estimates on a common scale across raters.

Second, the results from this study suggest that increasing the size of the link can improve the overall stability of rater severity estimates. This finding is important because one benefit of using MFR analysis is the ability to simultaneously evaluate rater performance along with examinee proficiency and task difficulty. Although the results from this study do not indicate a minimum number of common examinees that must be included in the linking set, and although prior research suggests that even one rating may be sufficient for linking purposes (Myford & Wolfe, 2000), increasing the number of shared ratings may provide a more precise and stable measure of rater severity. Along the same lines, larger sample sizes within the link have the potential to reduce the impact of idiosyncrasies in individual examinees within the link on the stability of rater severity estimates.

Finally, when using a block of shared ratings, care should be taken to include some examinees who are likely to elicit less consistent ratings. Somewhat counterintuitively, the results from this study and those reported by Myford and Wolfe (2000) suggested that selecting common performances with higher levels of misfit may result in higher levels of precision of parameter estimates. In operational settings, this could be accomplished by including some common examinees in the linking set who are known to be difficult for raters to rate consistently. This result can potentially be interpreted as somewhat akin to the effects of including a more discriminating item on a selected-response test. While these results do not suggest a specific ratio of noisy and acceptable fit within a linking set, we observed that adding only two such examinees to the link markedly improved the precision of the parameter estimates.

Limitations

When considering the results from this study, several limitations are important to note. First, we did not explore issues related to parameter recovery between the generating parameters for the simulation conditions and the estimated logit-scale locations that were obtained after introducing sparseness into the simulated data sets. Rather, we used the largest linking set sample size ($N_{\mathrm{Link}} = 8$) as a frame of reference from which to consider the stability of the estimates across modified linking set conditions. This analytic procedure reflects the nature of sparse rating designs in operational performance assessment settings, including teacher evaluation.

Sparse rating designs have the inherent limitation that the precision of the obtained parameter estimates will decrease as the sparseness of the rating design increases. In other words, the missingness in the data structure leads to inflated standard errors, which in turn lead to imprecision in parameter estimates (Eckes, 2015). While this study suggests that parameter estimates tend to be stable across changes in the number, model–data fit, and ability of the linked examinees, the results do not imply that increasing the size of a block of shared ratings in a sparsely connected design will increase the precision in any measurable way.

That being said, the purpose of this study was not to explore the parameter recovery of sparse rating designs. This topic is of importance, considering the prevalence of sparse designs in the field of education; however, little research has been conducted on how best to improve the precision of parameter estimates in such cases. More research should be undertaken to explore how the precision of parameter estimates can be improved when sparse rating designs are unavoidable. Such research may wish to empirically test the effect of various improvements to sparse rating designs, or compare the benefits of other forms of analysis besides Rasch, such as the generalizability theory, that may be appropriate (Sudweeks, Reeve, & Bradshaw, 2004).

Second, we did not manipulate or examine rater effects (e.g., leniency/severity, central tendency, or other types of range restriction, biases, etc.) in the simulated data. Accordingly, it was not possible to systematically explore the influence of these effects in the context of sparsely connected assessment systems. Particularly in assessment situations with only one rater per examinee, information regarding examinee proficiency may be inaccurate if the rater exhibits rating errors or systematic biases. Although raters who do show extreme rater effects can be identified through MFR analysis, linking the disconnected subsets in these situations would do little to circumvent the fact that ratings from these raters might be questionable. If possible, practitioners and researchers should weigh the feasibility of strengthening the design by including additional shared ratings between raters or use a rating design where all examinees are rated by multiple raters.

Another important limitation is related to the representativeness of the simulation design. Although the characteristics of the simulated data were intended to reflect a wide range of operational assessment contexts, other assessment contexts may differ in important ways from the simulated data explored in the current study; in some cases, these differences may limit the generalizability of the current results. In particular, the simulation design did not include systematic manipulation of the number of rating scale categories or tasks included in the analytic rubric. Furthermore, we did not manipulate characteristics related to examinee, rater, or task fit within the operational ratings, and we did not manipulate the ratio of examinees, raters, and tasks with acceptable and noisy fit within the operational or linking set ratings. Similarly, we did not systematically introduce or model dependencies that may result from nested assessment systems, such as teacher evaluation systems in which several teachers from the same school are evaluated by their own principal. Accordingly, it is not possible to draw conclusions about the generalizability of these findings to assessment contexts in which these characteristics are different from the simulation design or the real ratings.

Finally, although the results from this study appear promising with regard to the use of small linking sets to establish a common metric for calibrating examinees, raters, and tasks in sparsely connected teacher evaluation systems, the minimum sample size within the linking set needed to establish a psychometrically sound assessment system was not identified. As with all assessment procedures, the unique context of the assessment system, including the stakes associated with assessment-based decisions, should inform decisions regarding data collection designs and analytic approaches.

Directions for Future Research

The results from this study suggest several directions for future research. The first clear direction for future research is related to the stability of rater severity estimates across characteristics of the linking set. As noted above, the correlations between rater severity estimates were less stable across modified linking set sample sizes compared with the examinee and task facets. This result highlights the potential role of rater effects on the stability of rater severity estimates across data collection designs. Specifically, these results suggest that additional research is needed in which the influence of rater effects, such as severity/leniency and central tendency/extremism, within linking sets and operational ratings on estimates of examinee, rater, and tasks is systematically examined.

Similarly, additional research is needed in which the influence of examinee demographic characteristics within the linking set is explored, including the influence of the match between the demographic characteristics of examinees within the linking set and the operational ratings, as well as the match between the demographic characteristics of examinees and raters. Furthermore, additional simulation studies and real data analyses are needed to examine the influence of additional characteristics that were not included in the current study, including systematic explorations of different numbers of rating scale categories and tasks, varying levels of model–data fit within operational ratings, and nesting structures within operational ratings.

As noted above, the correlation analyses in this study were based on the correspondence between estimates obtained from the conditions in which the linking set included eight common examinees and the conditions with modified linking set sample sizes. This analytic approach reflects the operational teacher evaluation context that provided the substantive motivation for this study. Specifically, in large-scale teacher evaluation systems based on principal observations, fully crossed ratings are not available. Accordingly, the correspondence between the estimates based on conditions with the two modified linking set sample sizes and the original linking set sample size was more meaningful in the context of the current study than the correspondence between the observed parameter estimates and the generating parameters. However, examination of the estimates for the simulation conditions based on the original linking set sample size reveals a downward (negative) bias in examinee estimates, which were generated based on a normal distribution with a mean of zero logits (see Table 3). This result suggests that the removal of large proportions of ratings to arrive at the fully disconnected operational ratings resulted in some estimation bias when the ratings were analyzed using the Facets software. Additional research should include detailed examination of the influence of missing data on parameter recovery using simulation studies.

Appendix

This appendix illlustrates the procedure for introducing person misfit into the linking set in the simulated data using a small example based on a linking set with 5 raters, 10 examinees, and 1 task.

Step 1: Simulate linking set ratings following the design specifications. The generating rater parameters (λ) should match those used to simulate the operational ratings.

	Task 1				
Examinees	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$
2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$
3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$
4	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$	$X_{4,5}$
5	$X_{5,1}$	$X_{5,2}$	$X_{5,3}$	$X_{5,4}$	$X_{5,5}$
6	$X_{6,1}$	$X_{6,2}$	$X_{6,3}$	$X_{6,4}$	$X_{6,5}$
7	$X_{7,1}$	$X_{7,2}$	$X_{7,3}$	$X_{7,4}$	$X_{7,5}$
8	$X_{8,1}$	$X_{8,2}$	$X_{8,3}$	$X_{8,4}$	$X_{8,5}$
9	$X_{9,1}$	$X_{9,2}$	$X_{9,3}$	$X_{9,4}$	$X_{9,5}$
10	$X_{10,1}$	$X_{10,2}$	$X_{10,3}$	$X_{10,4}$	$X_{10,5}$

Note. The cell entries ($X_{i,j}$) reflect the rating (X) assigned to Examinee *i* by Rater *j*.

Step 2: Within each task, reorder the columns using random sampling without replacement to determine the new sequence of raters.

	Task 1				
Examinees	Rater 5	Rater 2	Rater 4	Rater 1	Rater 3
1	$X_{1,5}$	$X_{1,2}$	$X_{1,4}$	$X_{1,1}$	$X_{1,3}$
2	$X_{2,5}$	$X_{2,2}$	$X_{2,4}$	$X_{2,1}$	$X_{2,3}$
3	$X_{3,5}$	$X_{3,2}$	$X_{3,4}$	$X_{3,1}$	$X_{3,3}$
4	$X_{4,5}$	$X_{4,2}$	$X_{4,4}$	$X_{4,1}$	$X_{4,3}$
5	$X_{5,5}$	$X_{5,2}$	$X_{5,4}$	$X_{5,1}$	$X_{5,3}$
6	$X_{6,5}$	$X_{6,2}$	$X_{6,4}$	$X_{6,1}$	$X_{6,3}$
7	$X_{7,5}$	$X_{7,2}$	$X_{7,4}$	$X_{7,1}$	$X_{7,3}$
8	$X_{8,5}$	$X_{8,2}$	$X_{8,4}$	$X_{8,1}$	$X_{8,3}$
9	$X_{9,5}$	$X_{9,2}$	$X_{9,4}$	$X_{9,1}$	$X_{9,3}$
10	$X_{10,5}$	$X_{10,2}$	$X_{10,4}$	$X_{10,1}$	$X_{10,3}$

Step 3: Change the column labels to the order within the operational ratings, such that the contents of each column reflect discrepancies between the operational and linking sets.

	Task 1				
Examinees	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	$X_{1,5}$	$X_{1,2}$	$X_{1,4}$	$X_{1,1}$	$X_{1,3}$
2	$X_{2,5}$	$X_{2,2}$	$X_{2,4}$	$X_{2,1}$	$X_{2,3}$
3	$X_{3,5}$	$X_{3,2}$	$X_{3,4}$	$X_{3,1}$	$X_{3,3}$
4	$X_{4,5}$	$X_{4,2}$	$X_{4,4}$	$X_{4,1}$	$X_{4,3}$
5	$X_{5,5}$	$X_{5,2}$	$X_{5,4}$	$X_{5,1}$	$X_{5,3}$
6	$X_{6,5}$	$X_{6,2}$	$X_{6,4}$	$X_{6,1}$	$X_{6,3}$
7	$X_{7,5}$	$X_{7,2}$	$X_{7,4}$	$X_{7,1}$	$X_{7,3}$
8	$X_{8,5}$	$X_{8,2}$	$X_{8,4}$	$X_{8,1}$	$X_{8,3}$
9	$X_{9,5}$	$X_{9,2}$	$X_{9,4}$	$X_{9,1}$	$X_{9,3}$
10	$X_{10,5}$	$X_{10,2}$	$X_{10,4}$	$X_{10,1}$	$X_{10,3}$

Declaration of Conflicting Interests

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

Funding

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

References

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

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational and Behavioral Statistics*, *13*(1), 1-18. doi: 10.3102/10769986013001001

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement*, *26*, 321-338. doi: 10.1177/0146621602026003006

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*, 378-387. doi: 10.3102/0013189X16659442

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Cui, Y., & Mousavi, A. (2015). Explore the usefulness of person–fit analysis on large-scale assessment. *International Journal of Testing*, *15*, 23-49. doi:10.1080/15305058.2014.977444

DeMars, C. (2015). Estimating variance components from sparse data matrices in large-scale educational assessments. *Applied Measurement in Education*, *28*, 1-13. doi: 10.1080/08957347.2014.973562

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. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93-112. doi: 10.2307/1435170

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

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, *79*, 332-340. doi: 10.1037/0021-9010.79.3.332

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.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*, 486-512.

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, *19*, 171-200. doi:10.3102/10769986019003171

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227-238.

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*, i-34. doi:10.1002/j.2333-8504.2000.tb01832.x

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded ed., 1980, Chicago, IL: University of Chicago Press)

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, *30*, 253-268.

Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation & the Health Professions*, *14*, 100-122. doi: 10.1177/016327879101400107

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3-38.

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

Smith, R. M., Schumacker, R. E., & Bush, J. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261. doi:10.1016/j.asw.2004.11.001

Sykes, R. C., Ito, K., & Wang, Z. (2008). Rater effects and the assignment of raters to items. *Educational Measurement: Issues and Practices*, *27*, 44-45.

Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement*, *48*, 69-81. doi:10.1177/001316448804800109

Wind, S. A., Engelhard, G., Jr., & 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.

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

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, *8*, 204-234.