4-2023

Many-Facet Rasch Designs: How Should Raters be Assigned to Examinees?

Christine E. DeMars
demarsce@jmu.edu

Yelisey A. Shapovalov
James Madison University

John D. Hathcoat

Recommended Citation

DeMars, C. E., Shapovalov, Y. A., & Hathcoat, J. D. (2023, April 13-15). Many-Facet Rasch Designs: How Should Raters be Assigned to Examinees? [Paper presentation]. National Council on Measurement in Education annual meeting, Chicago.

Many-Facet Rasch Designs: How Should Raters be Assigned to Examinees?

Christine E. DeMars

Yelisey A. Shapovalov

John D. Hathcoat

Department of Graduate Psychology and

Center for Assessment and Research Studies,

James Madison University

Author Note

Correspondence concerning this manuscript should be addressed to Christine DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg VA 22807.

Many-Facet Rasch Designs: How Should Raters be Assigned to Examinees?

Abstract

In Facets models, raters should be connected, and there are multiple ways to connect raters. Keeping the number of ratings constant and two raters scoring each examinee, the standard error of both rater severity and examinee ability was higher when raters scored one examinee in common with many different raters than when they scored many examinees in common with two raters. However, the differences were small, especially for the standard error of examinee ability. Alternatively, when only a subset of examinees were scored by two or more raters, the smallest standard errors were achieved when all raters scored a common linking set, although standard errors were larger than they were when all examinees were scored by two raters. If the rating design only allows for a single rating of most examinees, it is preferable to link the metric by assigning all raters to rate the same set of linking examinees.

Many-facets Rasch models (MFRM) allow researchers to put measures of multiple facets, such as examinees, raters, tasks, or rating criteria, on the same metric **if** all of the levels of the facets are linked (Engelhard, 1997). However, if disjoint pairs of raters scored different subgroups of examinees, rater severity would be confounded with examinee ability. Linking is also sometimes called *connectivity*. Various designs can be used to link the raters, similar to designs used to link items for equating test forms. Links may be direct or indirect. When two raters rate the same responses, the link between the raters is termed *direct*. Rater 1 and rater 3 are indirectly linked when they rate no responses in common, but both raters have direct links with rater 2. Links can also be strong or weak. If two raters rate one response in common, they have a weak, direct link. If they rate many of the same responses, they have a strong link. In contexts where some or all examinees will be scored by two raters, possible designs include:

Design A (rotating pairs): Each rater scores one (or a few) examinee(s) in common with many different raters. For example, if each rater scores 50 examinees, each rater could be paired with 50 different raters if each rater pair only scores one examinee. Each rater has many weak direct links.

Design B (fixed pairs): Each rater scores one subset of examinees with one rater and another subset of examinees with another rater, linking all the raters in a chain. For example, rater 1 and rater 2 score a subset of examinees, rater 2 and rater 3 score another subset of examinees, etc. Rater 2 is directly linked to raters 1 and 3, and indirectly linked to other raters. Each rater has direct links with only two other raters, but they are strong links.

Design C (random): Raters are randomly assigned to examinees, approximating Method A.

If most examinees are only scored by one rater any of the above designs may be used for a subset of double-scored examinees. Additionally, a fourth design is common when most examinees are scored by a single rater:

Design D (linking subset): All raters score the same small subset of examinees to link the rater severities.

The focus of this study was on comparing the accuracy of estimates of rater severity and person ability using these designs. Is it better to have many direct but weak links, or a small number of strong direct links? Related studies of linking designs have not directly assessed these estimates. Wind (2022) showed that fixed-pairs designs (Design B) could be problematic for detecting differential rater functioning (DRF) but did not examine the precision of parameter estimates in the absence of DRF. Wang et al. (2021) compared a fixed-pair design to a design where most raters were disconnected but did not include other designs where the raters were linked. Guo and Wind (2023) compared the accuracy of rater severity estimates, but looked at designs in which a single rater scored each response and raters were linked through multiple-choice items or through a combination of multiple choice items and rotation of raters through different rater-scored items. Myford and Wolfe (2000) deleted data to form disconnected subsets and studied various linking subsets, but did not compare the results to the original designs which used some type of rater pairing, perhaps random. Using a small subset of examinees to anchor the ratings (Design D) but a single rater for most examinees, Wind and Jones (2018) found that the standard errors for raters, but not examinees, decreased as the number of examinees in the link increased. However, they did not include other linking designs for comparison.

Research Questions

1. Does the linking design impact the standard error of rater severity?
2. If the linking design impacts the estimates of rater severity, does that lead to differences in standard error of examinee ability?

In Study 1, each examinee was scored by two raters, as might be done in a high-stakes context. In Study 2, most examinees were scored by only one rater, but each rater scored some examinees in common with other raters, as might be done to link the raters or to collect data for inter-rater reliability estimation. In both studies, all raters were linked, either directly or indirectly; there were no disconnected subsets.

Study 1

Method

Data were simulated for 200 replications of 16 or 31 raters each scoring 30 or 60 examinees, with pairs of raters assigned using Designs A (rotating pairs), B (fixed pairs), and C (random pairs). As shown in Table 1, for Design A (rotating) each rater scored 1, 2, or 4 examinees in common with **each** of the other raters. Each examinee was scored by 2 raters on 5 tasks, using a 5-point rating scale.

The model for simulating item scores was:

$$\ln\left(\frac{P(x=k)}{P(x=k-1)}\right)=\beta_n-\lambda_r-\delta_i-\tau_k\ , \qquad (1)$$

where x is the observed score, β_n is ability for examinee *n*, λ_r is severity for rater *r*, δ_i is difficulty for task *i*, and τ_k is the difficulty of category *k* relative to category *k-1*.

Abilities were $\sim$.N(0,1) and severities were $\sim$.N(0,0.1^2), N(0,0.5^2) or N(0,0.8^2). To obtain the standard error for each ability or severity parameter, the same parameters were used in each replication, but the pairing of raters and the assignment of raters to examinees was randomized differently in each replication. The abilities and severities were generated from an inverse cumulative normal distribution with a mean of zero and the specified variance (for example, β_n is = Φ^{-1}(n/N), where N is the number of examinees and n is the rank for examinee n). Task difficulties (δ) were fixed with values (−0.6, −0.4, 0, 0.4, 0.6) and relative step offsets (τ) were (−1, −0.4, 0.4, 1).

The variance of the abilities was somewhat arbitrary; greater variance yields greater person separation reliability. The variance of the severities can be interpreted relative to the variance of the abilities, and rater training may reduce the variance in severities. Relative to the variance of the abilities, the variance of 0.1^2 was similar to some studies of writing (Gyagenda & Engelhard, 2010; Wind & Engelhard, 2012). The ratios of the rater variance to the ability variance in some studies of instrumental music performance (Wesolowski, 2019; Wind et al., 2016) were between the values of 0.5^2 and 0.8^2. The tasks could represent either different scoring elements (criteria, domains) for one examinee product, or they could represent different writing prompts or problem scenarios. One might expect a smaller range of difficulty for different writing prompts or different criteria used to evaluate the same task. But one might expect a larger range of difficulty for science tasks or math problems purposely selected to cover varying levels of difficulty. The range chosen here seemed a middle range of difficulty for performance tasks.

Parameters were estimated using Facets (Linacre, 2021). The SE was estimated for each ability as

$$\mathrm{SE}(\beta_n) = \sqrt{\sum\nolimits_j (\hat{\beta}_{nj} - \overline{\beta}_{n\cdot})^2 \big/ J} = \sqrt{\mathrm{MSE}_n - \mathrm{bias}_n^2}\ , \tag{2}$$

where $\hat{\beta}_{nj}$ is the estimate of ability for person n in replication j, $\overline{\beta}_{n\cdot}$ is the average estimate for person n, and J is the number of replications. Calculations were similar for rater severity.

To help put the SE of the ability estimates in context, reliability of the score estimates was calculated as the squared correlation between the true and estimated ability: $r^2(\hat{\beta}_n, \beta_n)$.

Results

For rater severity, the empirical SEs were nearly constant within the interval −1 to 1, so Figure 1 shows the mean empirical SEs (more precisely, the square root of the mean squared SE) for rater severity within that range.[1] The SE of course decreased as the number of ratings per rater increased. The SE increased as the standard deviation of rater severity increased. Relevant to the research question is the comparison of the linking methods: SEs were greater for Design B (fixed pairs) compared to Designs A and C. Design C (random) was mostly comparable to Design A (rotating pairs).

Figure 2 shows the mean SE for examinee ability (more precisely, the square root of the mean squared SE). The SE varied by ability, but the same abilities were used regardless of rater variance so taking the average did not distort comparisons. In the left panel, where the variance of rater severity was smallest, the rater design made almost no difference. In the middle and

[1] The SE was higher outside this range because there was less information for more extreme raters; the overall mean (not shown) increased somewhat for the condition with the greatest rater variance because there was a non-negligible proportion of more extreme raters.

right panels, where the variance in severity was greater, the SEs were somewhat greater for Design B (fixed pairs) compared to Designs A (rotating pairs) and C (random). In essence, the SE for rater severity contributed to the SE for the abilities—in some replications, an examinee might be assigned to raters whose severity was overestimated, and in other replications that examinee might be assigned to raters whose severity was underestimated, leading to greater instability. However, the difference between the SEs for Design B compared to Designs A and C was small.

The reliability of the score estimates provides one way to assess the practical impact of the examinee SEs. The reliability was defined as the squared correlation between the estimated and true ability. Figure 3 shows the reliability estimates for each condition. The difference between reliability for Designs A (rotating pairs) and Design B (fixed pairs) was never much more than .01, even when the rater SD was 0.8. Reliability decreased as the standard deviation of rater severity increased, and reliability was higher when there were more examinees per rater.

The difference in examinee SE or reliability from different rater linking designs would not be evident outside of a simulation study. With real data, the SEs are estimated analytically as the inverse of the information function. The errors in the estimates of the rater severities, task difficulties, etc. are not taken into account. Figure 4 shows the mean model-based SEs (square root of the mean squared SE) for examinee ability as reported in the software output. These SEs appear nearly identical for all rater designs. The analytical SEs in Figure 4 are slightly smaller than the empirical SEs in Figure 2. Thus, the SEs obtained analytically are deceptive in that it appears that the magnitude of the rater error does not impact the magnitude of the examinee SEs.

Although the SEs for examinee abilities were somewhat larger using Design B (fixed pairs) when the variance in rater severity was medium or large, it is also important to consider how much including rater severities in the model improved ability estimation compared to omitting rater from the model. Figure 5 shows the empirical reliability (squared correlation between true and estimated ability) of each rater-linking design including or omitting rater severity in the model. Any of the rating designs provided much more reliable scores than designs where the rater was omitted from the model, except when the rater SD was 0.1. Omitting the rater severity decreased the reliability by as much as 0.2.

Study 2

Method

Data were simulated for 200 replications of 16 or 31 raters each scoring 30 or 60 examinees, with pairs of raters assigned using Designs D (linking subset), B (fixed pairs), and C (random pairs). Each rater scored 4 or 8 examinees in common with other raters. As shown in Table 2, the total number of examinees scored varied with the rating design. Most examinees were scored by only one rater, but some examinees were scored by all raters (Design D) or by 2 raters (Designs B and C). Examinees were scored on 5 tasks, using a 5-point rating scale.

The model, distributions of abilities and rater severities, and values for item and category difficulties were the same as they were in Study 1.

Results

One replication with random pairs yielded disconnected subsets, as reported in the Facets output, so another dataset was generated to replace this replication.

Figure 6 shows the mean empirical SEs for rater severity within the range of −1 to 1 for the conditions where each rater scored 8 examinees in common with other raters and served as the sole rater for another 22 or 52 examinees. The pattern was similar for 4 common examinees (not shown), but the SEs were greater. The SEs were somewhat greater than in Study 1, where raters scored the same number of examinees but had more examinees in common. The SE decreased as the number of ratings per rater increased. The SE increased as the standard deviation of rater severity increased. Relevant to the research question is the comparison of the linking methods: SEs were smaller for Design D (linking subset) compared to Designs B and C (fixed or random pairs). Design B (fixed pairs) yielded somewhat smaller SEs than Design C (random pairs), except when the rater variance was large.

Figure 7 shows the SE for examinee ability for the examinees who were scored by a single rater. Design D (linking subsets) led to smaller SEs, as the more precise estimates of rater severity carried through to increase the precision of the abilities.

To assess the practical impact of the examinee SEs, the reliability of the score estimates is shown in Figure 8, again defined as the squared correlation between the estimated and true ability for examinees scored by a single rater. The difference between reliability for Designs D (linking subset) and Design B (fixed pairs) was never much more than .02. This difference was twice the largest differences in Study 1, but still seems small. Also, as in Study 1, omitting the rater severity parameter from the model yielded noticeably lower reliabilities (Figure 9). Any linking method was better than omitting rater severity from the model.

Conclusions and Implications

With single ratings for most examinees (Study 2), rater severity and examinee ability were estimated with smaller SE when all raters rated the same subset. With double-ratings (Study 1), rater severity and examinee ability were estimated with slightly smaller SE when each rater was directly linked to many other raters (with just a single examinee in each link) instead of direct links to two raters (with a larger group of examinees in the links). Given Wind's (2022) findings regarding detection of DRF, it is desirable to avoid repeatedly pairing each rater with a small number of raters and linking them in a chain if one wishes to assess DRF. Otherwise, if the concern is simply with the standard error of the abilities, random or systematically varying pairs (for double-ratings) are still slightly preferable, but fixed pairs would be acceptable if they are easier to implement. If the variance in rater severity is small, the linking design makes little difference but unless the assessment system is well-established one cannot know how much variance to expect.

It is important to note that all of the linking designs studied here connected all the raters. No designs involved disconnected subsets of raters. The conclusion that the linking design made little difference when examinees are double-scored should not be generalized to mean that connectivity does not matter.

References

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

Guo, W., & Wind, S. A. (2023). The effects of rating designs on rater classification accuracy and rater measurement precision in large-scale mixed-format assessments. *Applied Psychological Measurement*, *47*(2), 91-105. https://doi.org/10.1177/01466216231151705

Gyagenda, I. S., & Engelhard, G., Jr. (2010). Rater, domain, and gender influences on the assessed quality of student writing. In M. L. Garner, G. Engelhard, W. P. Fisher, & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 398-429). JAM Press.

Linacre, J. M. (2021). *Facets computer program for many-facet Rasch Measurement* (Version 3.83.5). Winsteps.com. http://www.winsteps.com/facets.htm

Myford, C. M., & Wolfe, E. W. (2000). Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs (Report No. RR-00-9). Educational Testing Service. https://dx.doi.org/10.1002/j.2333-8504.2000.tb01832.x

Wesolowski, B. C. (2019). Predicting operational rater-type classifications using Rasch measurement theory and random forests: A music performance assessment perspective. *Journal of Educational Measurement, 56*(3), 610-625. https://doi.org/10.1111/jedm.12227

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*(4) 278-299. https://dx.doi.org/10.1080/10627197.2016.1236676

Wind, S. A. (2022). Rater connections and the detection of bias in performance assessment. *Measurement: Interdisciplinary Research and Perspectives, 20*(2), 91-106. https://doi.org/10.1080/15366367.2021.1942672

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

Wind, S. A., & Jones, E. (2018). The stabilizing influences of linking set size and model-data fit in sparse rater-mediated assessment networks. *Educational and Psychological Measurement, 78*(4), 679-707. https://doi.org/10.1177/0013164417703733

Table 1

Rotating vs. Fixed Pairs

	Rotating						
	Rater						
Examinee	1	2	3	4	5	6	7
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	
11		x					x
12			x	x			
13			x		x		
14			x			x	
15			x				x
16				x	x		
17				x		x	
18				x			x
19					x	x	
20					x		x
21						x	x
	Fixed						
1 to 3	x	x					
4 to 6		x	x				
7 to 9			x	x			
10 to 12				x	x		
13 to 15					x	x	
16 to 18						x	x
19 to 21	x						x

Note. For brevity, the design shows 7 raters each rating 6 examinees, but can generalize to N raters each rating k(N-1) examinees.

Table 2

Number of Examinees and Rater Pairings, Study 1

	Total Examinees	Design A: Times paired with each other rater	Design B: Times paired with each of 2 other raters
16 Raters			
30 Examinees Scored by each Rater	240	2	15
60 Examinees Scored by each Rater	480	4	30
31 Raters			
30 Examinees Scored by each Rater	465	1	15
60 Examinees Scored by each Rater	930	2	30

Table 3

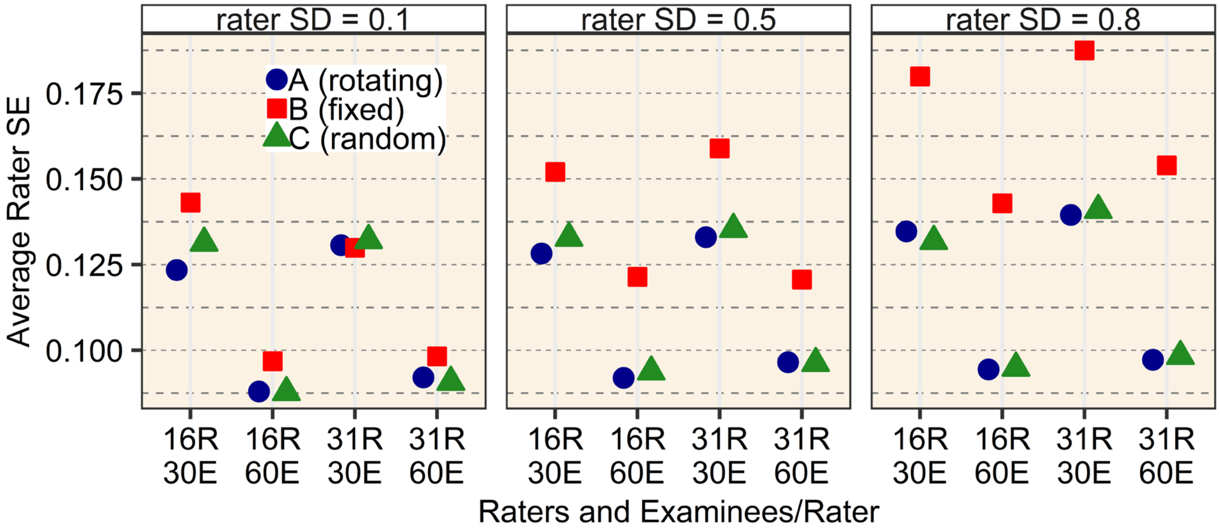
Number of Examinees and Rater Pairings, Study 2

	Design B or C		Design D	
	Total Examinees	Examinees with Single Rater	Total Examinees	Examinees with Single Rater
4 Common Ratings				
16 Raters				
30 Examinees Scored by each Rater	448	416	420	416
60 Examinees Scored by each Rater	928	896	900	806
31 Raters				
30 Examinees Scored by each Rater	868	806	810	806
60 Examinees Scored by each Rater	1798	1736	1740	1736
8 Common Ratings				
16 Raters				
30 Examinees Scored by each Rater	416	352	360	352
60 Examinees Scored by each Rater	896	832	840	832
31 Raters				
30 Examinees Scored by each Rater	806	682	690	682
60 Examinees Scored by each Rater	1736	1612	1620	1612

Note. In Design B, 4 common ratings indicates that each rater scored 2 examinees in common with each of 2 other raters. In Design C, 4 common ratings indicates that each rater scored 1 examinee in common with each of 4 other raters. In Design D, all raters scored the same linking set of 4 examinees.

Figure 1

Study 1: Mean Empirical Standard Error for Raters within -1 to 1 Logits



Note. 16R indicates 16 raters, 31R indicates 31 raters, 30E indicates each rater scored 30 examinees, 60E indicates each rater scored 60 examinees.

Figure 2

Study 1: Mean Empirical Standard Error for Examinee Ability

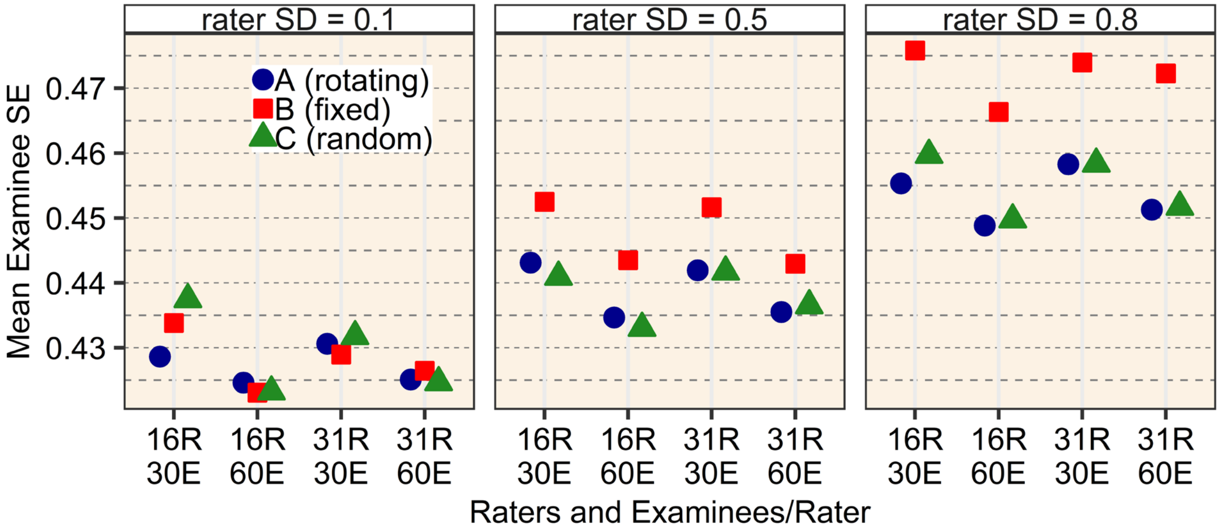


Figure 3

Study 1: Reliability Estimates

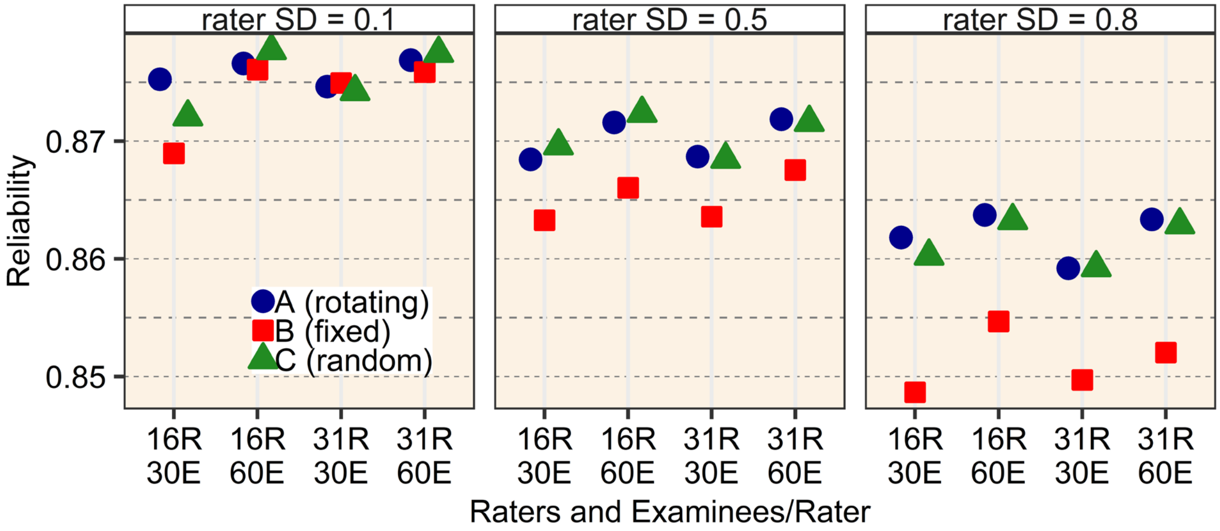
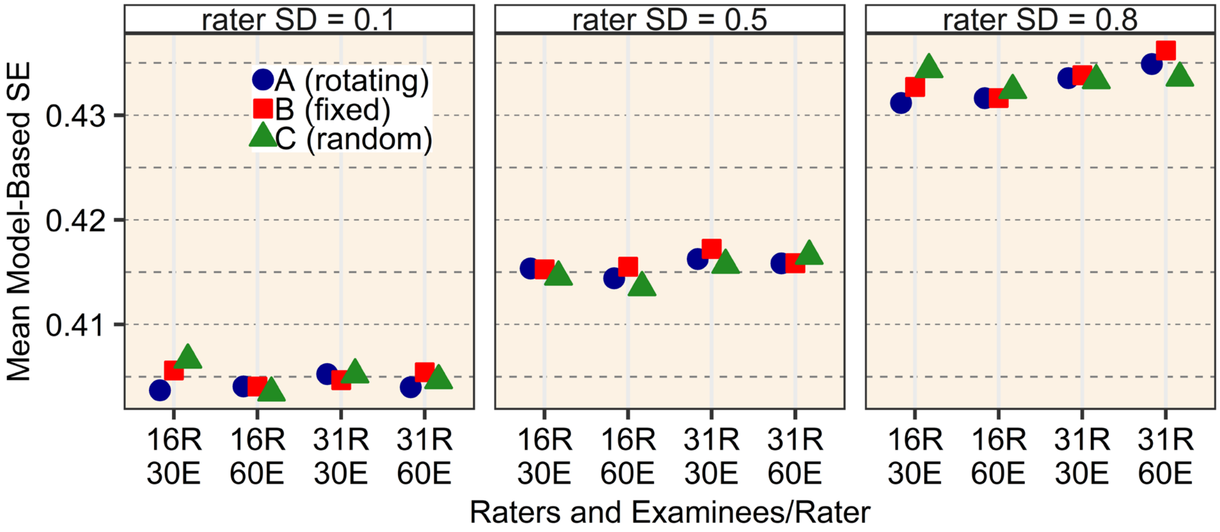


Figure 4

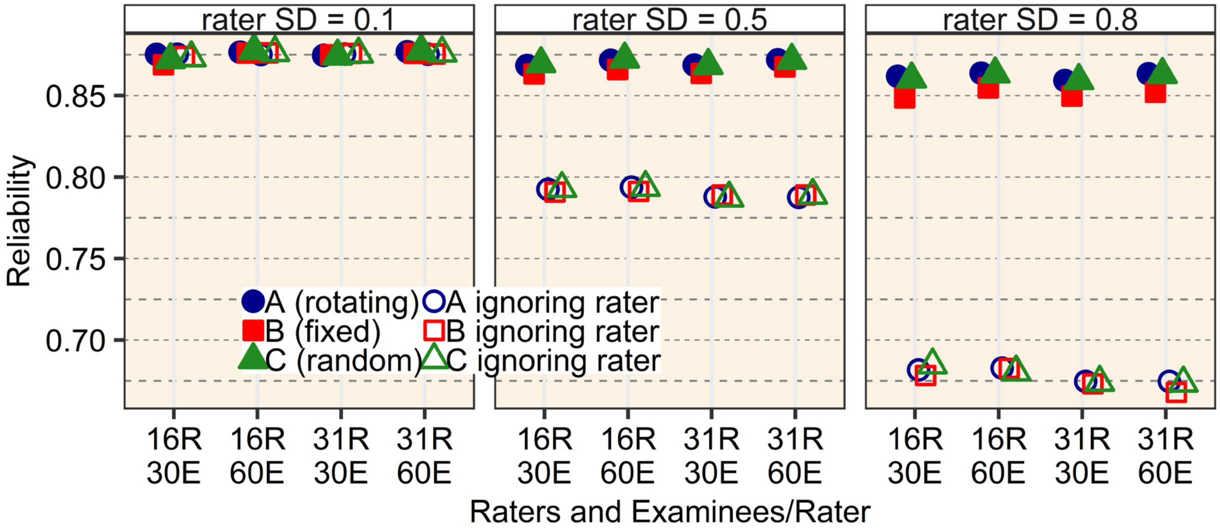
Study 1: Model-Based Standard Error for Examinee Ability



Note. The scale of the Y-axis is different than it was in Figure 2.

Figure 5

Study 1: Reliability Estimates, with and without Rater in the Model



Note. The scale of the Y-axis is different than it was in Figure 3 to accommodate the lower values of the models where the rater severity parameter was omitted.

Figure 6

Study 2: Mean Empirical Standard Error for Raters within -1 to 1 Logits

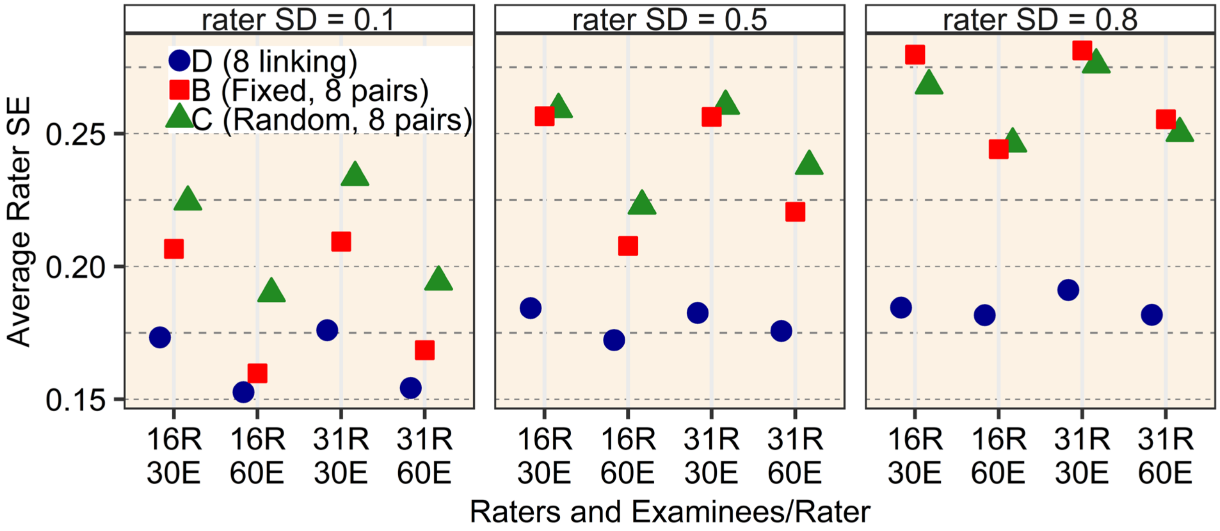


Figure 7

Study 2: Empirical Standard Error for Examinee Ability, Examinees Scored by only 1 Rater

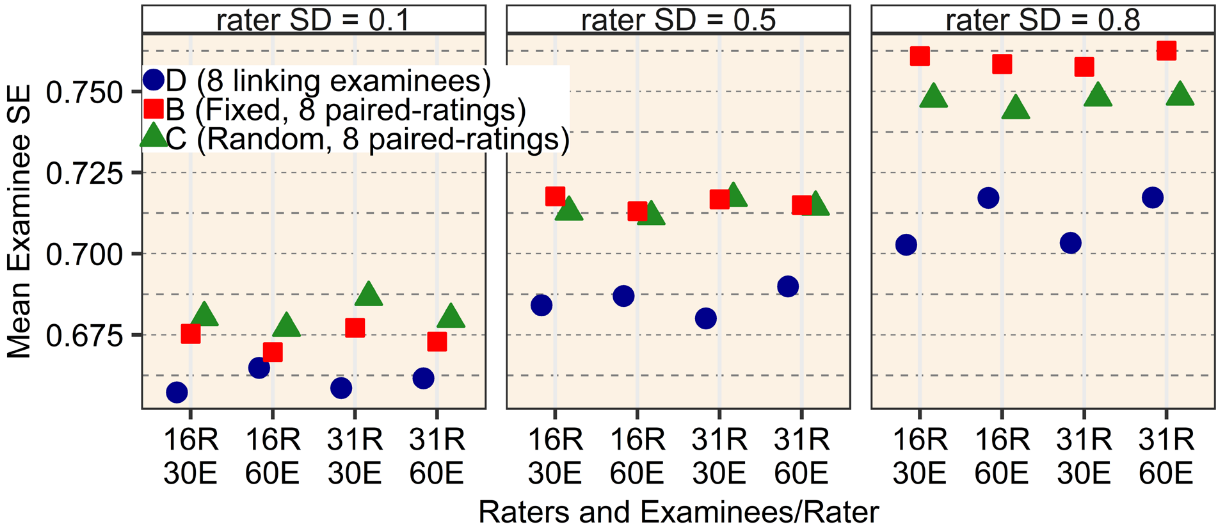


Figure 8

Study 2: Reliability Estimates for Examinees Scored by a Single Rater

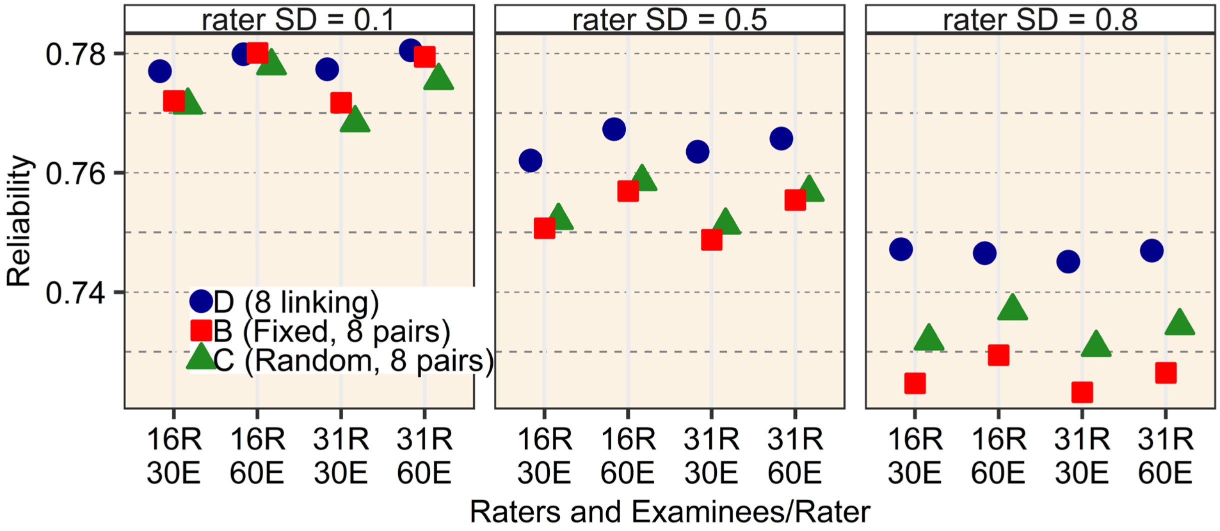
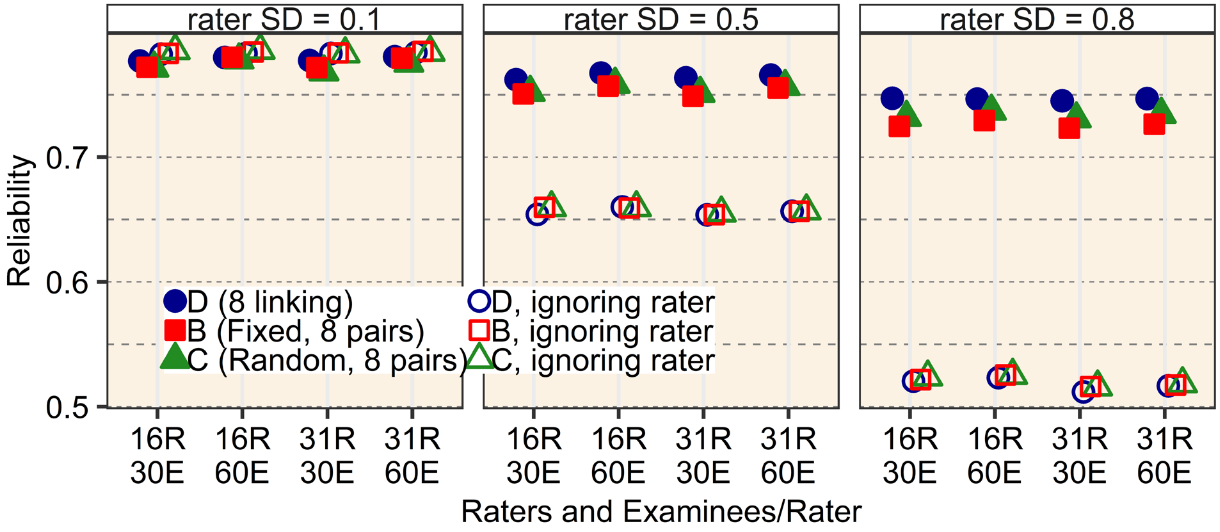


Figure 9

Study 2: Reliability Estimates, with and without Rater in the Model, for Examinees Scored by a Single Rater



Note. The scale of the Y-axis is different than it was in Figure 9 to accommodate the lower values of the models where the rater severity parameter was omitted.