A Comparison of Anchor Lengths and Item Selection Methods in Small-Sample Equating

Kylie N. Gorney

Department of Educational Psychology, University of Wisconsin-Madison

Author Note

Kylie N. Gorney https://orcid.org/0000-0002-8924-0726

Abstract

Two factors known to affect the quality of an equating anchor are length and item selection method. This simulation study considers the way in which both factors affect equating accuracy when samples are small. The following conditions were studied for a 100-item test: sample size ($N = 10, 25, 50, 100$), anchor length ($V = 20, 25, 30, 35, 40$), item selection method (minitest, semi-miditest, miditest), and equating method (identity, synthetic, nominal weights mean, circle-arc). Results showed that the minitest tended to produce the most accurate results, though the differences between item selection methods decreased as sample size increased. Additionally, the use of longer anchors generally improved results, especially for sample sizes of 10 and 25. Of the equating methods, synthetic equating was the most effective when the sample size was 10, while nominal weights mean equating or circle-arc equating were preferable for all other sample sizes.

Keywords: anchor test, small-sample equating

A Comparison of Anchor Lengths and Item Selection Methods in Small-Sample Equating

When samples are small, statistical procedures are often adjusted to account for the nature of the data. Test equating is no exception. In recent years, several equating methods have been compared in their ability to handle situations where there are fewer than 100 examinees per test form.

Perhaps the simplest method to consider is identity equating, which assumes that the two test forms being equated are completely parallel. Thus, an observed score on the new form is considered to be equivalent to the same observed score on the old form. Identity equating has been shown to perform well when samples are small and the test forms are similar in difficulty (e.g., Kim, von Davier, & Haberman, 2008; Skaggs, 2005). However, if the test forms differ even slightly in difficulty, then it is typically recommended that a different equating method be applied.

Nominal weights mean equating was specifically designed to handle small samples. It is a simplified version of Tucker mean equating (Gulliksen, 1950) that incorporates a ratio of the number of total test items to the number of anchor test items (Babcock, Albano, & Raymond, 2012). Nominal weights mean equating typically performs well when test forms differ in difficulty, though it is often bested by identity equating when the test forms are similar.

In practice, the true difference between test form difficulties is unknown. Therefore, it may be useful to consider some sort of compromise between these two methods. A synthetic linking function offers one such solution as it computes a weighted average of the identity function and a second equating function (Kim et al., 2008). The performance of this method largely depends on the data and the weights that are applied (Babcock et al., 2012; Kim, von Davier, & Haberman, 2011). Generally, it is useful to increase the weight on the identity function when the test forms are believed to be similar in difficulty, but

decrease this weight as the sample size increases. Alternatively, weights may be chosen such that the symmetry property of equating is maintained (Holland & Strawderman, 2011).

Still, another idea entirely is to relax one or more of the model assumptions. Identity equating, nominal weights mean equating, and synthetic equating all assume that the equating relationship is linear. However, circle-arc equating allows for the use of an equating curve and thus, in principle, considers a wider variety of situations (Livingston & Kim, 2009). This curve connects two prespecified endpoints and one empirically determined middle point. The middle point is determined by equating at one point in the middle of the score distribution. Moreover, any of the aforementioned methods may be used as the equating function. As was the case with synthetic equating, the results of circle-arc equating vary depending on a number of factors (e.g., the data, the prespecified endpoints, the equating function). However, when circle-arc equating has been used in small-sample situations, it has generally produced favorable results (Babcock et al., 2012; Kim & Livingston, 2010; Livingston & Kim, 2009).

With the exception of identity equating, all of the equating methods mentioned thus far rely on a set of common anchor items that are used in the computation. As such, it is sensible to consider the way in which these anchor items are selected. It is typically recommended that anchor items be representative in content and that they maintain certain statistical properties (Kolen & Brennan, 2014). Specifically, the mean and the spread of the anchor item difficulties are often constrained to be similar to those of the test forms being equated. This type of anchor is often referred to as a minitest.

In some situations, however, it may be challenging to construct an anchor that satisfies all of these requirements. Consider that in order to mimic the spread of item difficulties of the total test, a minitest may need to include items that are very easy or very difficult. In practice, such items are rare since they are often discarded for having low discriminatory power. Thus, Sinharay and Holland (2006) proposed two methods that relax the requirement that anchor items must have a spread of difficulties similar to that of

the total test. The first method, the miditest, selects items with a very small spread of difficulties. More specifically, the anchor items are selected to have difficulties that are similar to the average difficulty of the total test. The advantage of this method is that there should be a wide array of items to choose from since moderately difficult items are unlikely to be discarded. The second method, the semi-miditest, offers a compromise between the minitest and the miditest. That is, it allows the spread of the anchor item difficulties to be less than that of the total test while maintaining that it is more than that of a miditest.

Both the miditest and the semi-miditest have been shown to produce equating results comparable to those of minitests when an external anchor is used (e.g., Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011; Liu, Sinharay, Holland, Feigenbaum, & Curley, 2011; Sinharay & Holland, 2007). However, few have focused on situations where an internal anchor is used or where the sample size is smaller than 100. One exception is a study that was conducted by Furter and Dwyer (2020) in which 10 examinees were simulated to take each test form. They considered several item selection methods, including what they referred to as the unrestricted p-values method and the restricted p-values method. These methods roughly correspond to a minitest and semi-miditest, respectively, and both were found to produce similar results when nominal weights mean equating was used as the equating method.

Another aspect to consider regarding the quality of an anchor is the number of anchor items that are selected. Generally, the use of more anchor items reduces the random equating error (Budescu, 1985). However, including more items is less desirable from a test security standpoint due to the risk of item compromise. Thus, the optimal anchor length is likely to vary by testing program, though a general rule of thumb is that an anchor should be at least 20% of the length of a test containing 40 or more items, or at least 30 items if the test is very long (Kolen & Brennan, 2014).

The purpose of this study is to compare the effects of using different anchor lengths and item selection methods in small-sample equating. Though prior research has examined

each of these factors separately, none have crossed them. By doing so, this study considers the possibility that different item selection methods may have different anchor length requirements. Additionally, because the true differences in test form difficulties and group abilities are unknown in practice, they are treated as random effects, thereby allowing for a wide range of plausible scenarios. This differs from previous research, where these differences have been fixed. Finally, all of the conditions considered here are examined using small samples, specifically. As such, the results of this study may prove useful for the many small-volume testing programs that seek to maximize equating accuracy.

Method

Data

The data used in this study were simulated using the three-parameter logistic model in R (Birnbaum, 1968; R Core Team, 2021). This model was chosen because it has been shown to accurately model the data of several testing programs (e.g., Babcock et al., 2012). Both the old and new test forms were simulated to contain 100 total items. Five anchor lengths ($V = 20, 25, 30, 35, 40$) and four sample sizes ($N = 10, 25, 50, 100$) were considered. Additionally, 1,000 replications were conducted for each condition.

For both the old and new test forms, the a parameters were sampled from a lognormal distribution (0,0.1), and the c parameters were set equal to 0.2. The b parameters of the old form followed $N(0, 1)$, while the b parameters of the new form were sampled from $N(\Delta_b, 1)$ where Δ_b represents the difference in test form difficulties. In practice, the true difference between test form difficulties is unknown; thus, the Δ_b parameters were sampled from $N(0, 0.25)$ to allow for instances where the new form was more difficult than the old form, and vice versa. The ability parameters of the old group followed $N(0, 1)$, while the ability parameters of the new group were sampled from $N(\Delta_\theta, 1)$ where Δ_θ represents the difference in group abilities. In particular, the Δ_θ parameters were sampled from $N(0, 0.25)$.

Anchor Item Selection Methods

After simulating the old form data, anchor items were randomly selected according to one of three methods. Note that the term *p-value* in this section refers to the proportion of examinees who answered an item correctly. This statistic was used as the measure of item difficulty rather than item response theory difficulty parameter estimates because of the small samples that were used.

All item selection methods used the following criteria: (1) the anchor item p-values could not be 0 or 1 (i.e., these items could not have been answered incorrectly by all of the examinees in the old group, nor could they have been answered correctly by all of the examinees in the old group), (2) the mean of the anchor test p-values had to be within 0.01 of the mean of the old form p-values, and (3) the standard deviation of the anchor test p-values had to be within 0.01 of a certain percentage of the standard deviation of the old form p-values. These percentages were 100%, 90%, and 75%, and they were designed to correspond to a minitest and two semi-miditests, respectively. For simplicity, the semi-miditest with the 75% spread is referred to as a miditest from this point forward.

There were two instances where all three requirements could not be satisfied. Both instances occurred when the sample size was 10 and a minitest was used. In these cases, the standard deviation requirement was relaxed. That is, the standard deviation of the anchor test p-values was only required to be within 0.02 of the standard deviation of the old form p-values.

Equating Design and Methods

Four equating methods were evaluated (Gorney, 2021), each of which is able to be applied under a common-item nonequivalent groups design. The first method, identity equating, was used as a baseline measure against which all other methods could be compared. If identity equating produced the smallest amounts of error, then it could be said that *not* equating actually led to the most accurate results. The second method that

was considered was nominal weights mean equating, as it has been shown to perform well in small-sample situations. The third method, synthetic equating, used a weighted average of the identity function and the nominal weights mean equating function. Because so many different conditions were considered, weights were chosen such that the symmetry property of equating was maintained. Specifically, the following equation was used to compute the weight placed on the nominal weights mean equating function:

$$w_e = \frac{(1 + b_e^2)^{-1/2}}{(1 + b_e^2)^{-1/2} + (1 + b_{id}^2)^{-1/2}} \tag{1}$$

where b_e is the slope of the nominal weights mean equating function, and b_{id} is the slope of the identity function (Holland & Strawderman, 2011). The fourth method, circle-arc equating, used nominal weights mean equating to determine the middle point. Additionally, 20 and 100 were used as the prespecified endpoints since 20 corresponds to the number of items an examinee of low ability was expected to answer correctly by guessing, and 100 corresponds to the number of total items on each test form. These values follow the recommendation given by Livingston and Kim (2009).

Evaluation Criteria

In order to compare the equated scores across replications, two criteria were assessed: bias and the root mean squared error (RMSE). Bias measures whether the estimated equated scores tend to over- or underestimate the true scores. In this study, an examinee's true score was defined as the one they received after being simulated to take the old form. The bias of an equating function may be written as

$$\text{Bias} = \frac{1}{N} \sum_{j=1}^{N} (\hat{y}_j - y_j) \tag{2}$$

where j denotes an examinee, N is the total number of examinees, $\hat{y}_j$ is an estimated equated score, and y_j is the true score. In contrast, the RMSE is concerned with the absolute difference between an estimated equated score and the true score. Thus, the

direction of error is less important than the fact that the error exists at all.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j - y_j)^2} \tag{3}$$

For each replication, the bias and RMSE were compared across sample sizes, item selection methods, anchor lengths, and equating methods. The conditions that used the most effective methods should yield values closer to 0.

Results

For each replication, the difference in test form difficulties changed, as did the difference in group abilities. Both sets of differences were simulated to follow a normal distribution with mean 0 and variance 0.25 to encompass a variety of situations that could be encountered in practice. The actual distributions of these differences may be viewed in Figure 1.

Tables 1–4 present the means and standard errors of the bias and RMSE for each condition. Figures 2 and 3 summarize these results by plotting the average bias and RMSE of each equating method against the anchor length. For consistency, both figures use one row to represent each sample size and one column to represent each item selection method. Moreover, each plot abbreviates the four equating methods as follows: identity equating (Id), synthetic equating (Syn), nominal weights mean equating (NM), and circle-arc equating (Cir).

Anchor Lengths and Item Selection Methods

Across all conditions, the minitest tended to produce the least biased results while the miditest tended to produce the most biased results (Figure 2). The semi-miditest generally fell somewhere in between. This held true for all of the sample sizes that were considered, though any differences between item selection methods became less pronounced as sample size increased. For each of the item selection methods, anchor length appeared

to have little to no impact on the average bias. In other words, an anchor with 40 items tended to exhibit just as much bias as an anchor with 20 items. However, the use of longer anchors tended to produce more stable results, and this was especially noticeable when the sample size was 10 or 25.

With respect to the RMSE, the results varied based on sample size (Figure 3). When the sample size was 10, the minitest produced slightly smaller RMSEs, on average, compared to the semi-miditest or the miditest. However, for sample sizes of 25 or more, the three item selection methods performed very similarly, with none offering a clear advantage over the others. Moreover, all of the item selection methods tended to yield smaller RMSEs as anchor length increased. This held true across all sample sizes, though it was particularly noticeable for the smaller sample sizes. In addition, when the sample size was 10, an increase in anchor length also corresponded to a decrease in RMSE variability. In other words, longer anchors tended to produce more consistent RMSEs than shorter anchors when the sample size was very small.

Equating Methods

Across nearly all conditions, identity equating yielded the most inconsistent amounts of bias (Table 1). Synthetic equating yielded relatively small amounts of bias, and it tended to be the most stable equating method when the sample size was 10 (Table 2). Nominal weights mean equating and circle-arc equating produced more biased results, on average, and this was especially noticeable when a miditest or a semi-miditest was used (Table 3; Table 4). Overall, the differences between equating methods were consistent across anchor lengths. That is, the difference between the bias of identity equating and the bias of circle-arc equating was roughly the same when 20 anchor items were used and when 40 anchor items were used.

Identity equating performed quite poorly with respect to the RMSE. Across all sample sizes and item selection methods, identity equating produced the largest and most

inconsistent RMSEs of any equating method. In contrast, synthetic equating performed very well when the sample size was 10, though it suffered across all other sample sizes. In these cases, nominal weights mean equating and circle-arc equating typically produced the smallest and most stable RMSEs. In fact, nominal weights mean equating and circle-arc equating performed so similarly in this regard that it is nearly impossible to say which method was superior.

Discussion

Several equating methods have been shown to perform well when samples are small, including identity equating, synthetic equating, nominal weights mean equating, and circle-arc equating. For three of these methods, anchor items play a central role in determining the quality of the equating. Thus, it is pertinent to consider the way in which these items are selected.

Three common item selection methods are the minitest, the semi-miditest, and the miditest. Each uses a different rule to determine the spread of the anchor item difficulties, and all have been shown to be effective when sample sizes are large. However, few studies have compared these item selection methods in situations where there are fewer than 100 examinees per test form, and none have considered the way in which anchor length could affect the results, as well. Therefore, the purpose of this study was to examine how anchor length and item selection method affect equating results when samples are small. In order to encompass a variety of situations, 1,000 replications were conducted in which the difference in test form difficulties and the difference in group abilities were treated as random.

Of the three item selection methods, the minitest tended to produce the most accurate results while the miditest tended to produce the least accurate results. As sample size increased, the differences between methods decreased, though they were still present for sample sizes of 100. An increase in anchor length typically decreased the variability of

the results, especially when smaller sample sizes were used. The use of longer anchors did not seem to affect the average bias, though it did correspond to a decrease in the average RMSE. Again, these effects were most noticeable when sample sizes were small, though they persisted for larger sample sizes, as well.

Identity equating generally produced less biased results, on average, though this benefit came with a cost: unstable results and large RMSEs. Synthetic equating performed well when the sample size was 10, but it yielded relatively large RMSEs for the larger sample sizes. Nominal weights mean equating and circle-arc equating performed better in this regard, as they produced the smallest RMSEs when the sample size was 25, 50, or 100. However, they also yielded the largest amounts of bias, on average, though this generally decreased as sample size increased.

As is often the case, this study was affected by a series of limitations. First, the two groups of examinees were simulated to be equal in size. In practice, one group is likely to be larger than the other, and this has the potential to affect the equating relationship. Second, though the differences in test form difficulties and group abilities were considered to be random, they did, in fact, follow a particular distribution. If a different distribution had been selected, say a normal distribution with a mean other than 0, then the overall conclusions would likely be affected. Third, three very specific item selection methods were considered. Though they have been labelled as a minitest, a semi-miditest, and a miditest, there exist several ways in which these types of anchors could be constructed. For example, a semi-miditest could be defined as one in which the standard deviation of the anchor item p-values is constrained to be 50% of the standard deviation of the old form p-values. Or, a different measure of item difficulty could be used altogether, such as item response theory difficulty parameter estimates. The fourth limitation is that content representation was not considered when selecting the anchor items. If content restrictions had also been imposed, then the anchor items may not have been able to fully satisfy the statistical requirements. Thus, this study only considered the case where content representation was not a concern.

The fifth limitation is that three of the four equating methods were based on nominal weights mean equating. Though this works well for drawing comparisons, it would be beneficial to consider different equating functions.

Future research is needed to address each one of these limitations. It would be useful to conduct additional simulation studies that examine different conditions, and real data examples are needed, as well. Any findings would be helpful in determining best practices for small-sample equating.

References

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement*, *72*(4), 608–628. https://doi.org/10.1177/0013164411428609

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, *22*(1), 13–20.

Furter, R. T., & Dwyer, A. C. (2020). Investigating the classification accuracy of Rasch and nominal weights mean equating with very small samples. *Applied Measurement in Education*, *33*(1), 44–53. https://doi.org/10.1080/08957347.2019.1674307

Gorney, K. N. (2021). *eqt: Small-sample equating* (Version 0.0.0.9000) [Computer software]. https://github.com/kyliegorney/eqt

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Holland, P. W., & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), *Statistical methods for test equating, scaling, and linking* (pp. 89–107). Springer. https://doi.org/10.1007/978-0-387-98138-3_6

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*(3), 286–298.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45*(4), 325–342.

Kim, S., von Davier, A. A., & Haberman, S. (2011). Practical application of a synthetic linking function on small-sample equating. *Applied Measurement in Education*, *24*, 95–114. https://doi.org/10.1080/08957347.2011.554601

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, *46*(3), 330–343.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Journal of Educational Measurement*, *48*(4), 361–379.

Liu, J., Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement*, *71*(2), 346–361. https://doi.org/10.1177/0013164410375571

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.0) [Computer software]. https://www.R-project.org/

Sinharay, S., & Holland, P. (2006). *The correlation between the scores of a test and an anchor test* (RR-06-04). Educational Testing Service.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*(3), 249–275.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, *42*(4), 309–330.

Table 1

Errors for Identity Equating (Means and Standard Errors)

V	Mini	Semi	Midi
$N = 10$			
20	0.0 (4.5)	0.0 (4.4)	0.1 (4.5)
25	0.0 (4.4)	0.0 (4.4)	0.0 (4.4)
30	−0.1 (4.3)	−0.1 (4.4)	0.0 (4.4)
35	0.0 (4.4)	−0.1 (4.4)	0.0 (4.4)
40	0.0 (4.3)	0.0 (4.4)	0.0 (4.3)
$N = 25$			
20	0.0 (4.2)	0.1 (4.2)	0.0 (4.2)
25	0.0 (4.2)	0.0 (4.1)	0.0 (4.1)
30	0.1 (4.2)	0.2 (4.2)	0.0 (4.1)
35	0.0 (4.1)	0.1 (4.1)	0.0 (4.1)
40	0.1 (4.1)	0.1 (4.1)	0.1 (4.0)
$N = 50$			
20	0.1 (3.9)	0.1 (4.0)	0.1 (4.1)
25	0.1 (4.0)	0.1 (4.0)	0.1 (3.9)
30	0.1 (4.0)	0.0 (3.9)	0.1 (4.0)
35	0.1 (4.0)	0.1 (3.9)	0.1 (3.9)
40	0.2 (3.9)	0.1 (3.9)	0.1 (3.9)
$N = 100$			
20	0.1 (3.9)	0.0 (3.9)	0.1 (4.0)
25	0.0 (4.0)	0.0 (4.0)	0.1 (3.9)
30	0.1 (4.0)	0.0 (3.9)	0.0 (4.0)
35	0.1 (3.9)	0.1 (3.8)	0.0 (3.9)
40	0.0 (4.0)	0.0 (3.9)	0.1 (3.9)

(a) *Bias*

V	Mini	Semi	Midi
$N = 10$			
20	7.2 (2.0)	7.1 (2.0)	7.2 (2.0)
25	7.2 (2.0)	7.2 (2.0)	7.2 (2.0)
30	7.1 (2.0)	7.2 (2.0)	7.1 (2.0)
35	7.1 (2.0)	7.1 (2.0)	7.2 (1.9)
40	7.1 (1.9)	7.2 (2.0)	7.1 (2.0)
$N = 25$			
20	7.3 (1.7)	7.3 (1.7)	7.3 (1.7)
25	7.3 (1.7)	7.3 (1.7)	7.2 (1.7)
30	7.3 (1.7)	7.3 (1.7)	7.2 (1.6)
35	7.2 (1.6)	7.2 (1.6)	7.2 (1.6)
40	7.2 (1.6)	7.2 (1.6)	7.2 (1.5)
$N = 50$			
20	7.2 (1.5)	7.3 (1.5)	7.3 (1.5)
25	7.3 (1.5)	7.2 (1.5)	7.2 (1.4)
30	7.3 (1.5)	7.2 (1.4)	7.2 (1.4)
35	7.3 (1.5)	7.2 (1.4)	7.2 (1.4)
40	7.2 (1.4)	7.2 (1.4)	7.2 (1.4)
$N = 100$			
20	7.2 (1.4)	7.2 (1.3)	7.3 (1.4)
25	7.3 (1.3)	7.3 (1.4)	7.3 (1.4)
30	7.3 (1.4)	7.2 (1.3)	7.3 (1.4)
35	7.3 (1.3)	7.2 (1.3)	7.3 (1.4)
40	7.3 (1.3)	7.2 (1.3)	7.2 (1.3)

(b) *RMSE*

Table 2

Errors for Synthetic Equating (Means and Standard Errors)

V	Mini	Semi	Midi
$N = 10$			
20	0.0 (3.3)	0.0 (3.3)	−0.1 (3.3)
25	0.0 (3.1)	−0.1 (3.2)	0.0 (3.2)
30	−0.1 (3.0)	−0.1 (3.2)	−0.2 (3.1)
35	0.0 (3.0)	−0.2 (3.1)	−0.2 (3.0)
40	0.0 (2.9)	0.0 (2.9)	−0.1 (3.0)
$N = 25$			
20	0.0 (2.6)	0.0 (2.7)	−0.2 (2.7)
25	0.0 (2.6)	−0.1 (2.5)	−0.1 (2.5)
30	0.0 (2.5)	0.0 (2.5)	−0.1 (2.5)
35	0.0 (2.4)	0.0 (2.5)	−0.1 (2.5)
40	0.0 (2.4)	0.0 (2.4)	−0.1 (2.4)
$N = 50$			
20	0.0 (2.3)	0.0 (2.3)	0.0 (2.3)
25	0.1 (2.3)	0.0 (2.2)	0.0 (2.2)
30	0.0 (2.2)	0.0 (2.2)	0.0 (2.2)
35	0.0 (2.2)	0.0 (2.1)	0.0 (2.2)
40	0.1 (2.2)	0.0 (2.2)	0.0 (2.2)
$N = 100$			
20	0.0 (2.1)	0.0 (2.1)	0.0 (2.2)
25	0.0 (2.1)	0.0 (2.2)	0.0 (2.1)
30	0.1 (2.1)	0.0 (2.1)	−0.1 (2.1)
35	0.1 (2.1)	0.0 (2.0)	0.0 (2.1)
40	0.0 (2.1)	0.0 (2.0)	0.0 (2.0)

(a) *Bias*

V	Mini	Semi	Midi
$N = 10$			
20	6.6 (1.6)	6.6 (1.6)	6.7 (1.6)
25	6.6 (1.6)	6.6 (1.6)	6.6 (1.6)
30	6.5 (1.6)	6.6 (1.6)	6.5 (1.6)
35	6.5 (1.6)	6.6 (1.6)	6.6 (1.5)
40	6.5 (1.5)	6.6 (1.6)	6.5 (1.6)
$N = 25$			
20	6.6 (1.1)	6.6 (1.1)	6.7 (1.2)
25	6.6 (1.1)	6.6 (1.1)	6.6 (1.1)
30	6.6 (1.1)	6.6 (1.1)	6.6 (1.1)
35	6.5 (1.1)	6.6 (1.1)	6.6 (1.1)
40	6.5 (1.0)	6.6 (1.1)	6.5 (1.0)
$N = 50$			
20	6.6 (0.8)	6.6 (0.8)	6.6 (0.8)
25	6.6 (0.8)	6.6 (0.8)	6.6 (0.8)
30	6.6 (0.8)	6.6 (0.8)	6.5 (0.8)
35	6.6 (0.8)	6.5 (0.8)	6.6 (0.8)
40	6.5 (0.8)	6.5 (0.8)	6.5 (0.8)
$N = 100$			
20	6.5 (0.7)	6.6 (0.6)	6.6 (0.7)
25	6.6 (0.6)	6.6 (0.7)	6.6 (0.7)
30	6.6 (0.7)	6.5 (0.6)	6.6 (0.7)
35	6.6 (0.6)	6.5 (0.6)	6.5 (0.7)
40	6.5 (0.7)	6.5 (0.7)	6.6 (0.6)

(b) *RMSE*

Table 3

Errors for Nominal Weights Mean Equating (Means and Standard Errors)

V	Mini	Semi	Midi
$N = 10$			
20	0.0 (4.1)	0.0 (4.1)	−0.2 (4.1)
25	0.0 (3.5)	−0.3 (3.5)	0.0 (3.6)
30	−0.1 (3.3)	−0.2 (3.5)	−0.3 (3.3)
35	0.0 (3.1)	−0.3 (3.1)	−0.3 (3.1)
40	0.0 (3.0)	−0.1 (2.8)	−0.3 (3.0)
$N = 25$			
20	−0.1 (2.7)	0.0 (2.7)	−0.3 (2.7)
25	−0.1 (2.4)	−0.2 (2.4)	−0.2 (2.4)
30	0.0 (2.3)	−0.1 (2.1)	−0.2 (2.2)
35	0.0 (2.0)	−0.1 (2.0)	−0.2 (2.0)
40	0.0 (1.9)	−0.1 (1.9)	−0.2 (1.8)
$N = 50$			
20	−0.1 (1.9)	−0.1 (2.0)	−0.1 (2.0)
25	0.1 (1.7)	0.0 (1.8)	−0.1 (1.7)
30	0.0 (1.6)	−0.1 (1.6)	−0.1 (1.6)
35	0.0 (1.5)	−0.1 (1.4)	−0.1 (1.5)
40	0.1 (1.4)	0.0 (1.4)	−0.1 (1.4)
$N = 100$			
20	0.0 (1.4)	−0.1 (1.4)	−0.1 (1.4)
25	0.0 (1.2)	−0.1 (1.2)	−0.1 (1.3)
30	0.0 (1.1)	−0.1 (1.1)	−0.2 (1.2)
35	0.0 (1.0)	−0.1 (1.0)	−0.1 (1.1)
40	0.0 (1.0)	−0.1 (1.0)	−0.1 (1.0)

(a) *Bias*

V	Mini	Semi	Midi
$N = 10$			
20	7.0 (1.8)	7.0 (1.8)	7.0 (1.8)
25	6.8 (1.7)	6.7 (1.7)	6.7 (1.7)
30	6.6 (1.7)	6.8 (1.7)	6.6 (1.6)
35	6.6 (1.6)	6.6 (1.6)	6.6 (1.5)
40	6.5 (1.6)	6.5 (1.5)	6.5 (1.5)
$N = 25$			
20	6.6 (1.1)	6.6 (1.1)	6.7 (1.1)
25	6.5 (1.0)	6.6 (1.0)	6.5 (1.1)
30	6.5 (1.1)	6.5 (1.0)	6.5 (1.0)
35	6.4 (1.0)	6.4 (1.0)	6.4 (1.0)
40	6.4 (0.9)	6.4 (0.9)	6.3 (1.0)
$N = 50$			
20	6.5 (0.7)	6.5 (0.8)	6.5 (0.7)
25	6.4 (0.7)	6.4 (0.7)	6.4 (0.7)
30	6.4 (0.7)	6.4 (0.7)	6.4 (0.7)
35	6.4 (0.7)	6.4 (0.6)	6.4 (0.7)
40	6.3 (0.7)	6.3 (0.7)	6.3 (0.7)
$N = 100$			
20	6.4 (0.5)	6.4 (0.5)	6.4 (0.5)
25	6.4 (0.5)	6.4 (0.5)	6.4 (0.5)
30	6.3 (0.5)	6.3 (0.5)	6.4 (0.5)
35	6.3 (0.5)	6.3 (0.5)	6.3 (0.5)
40	6.3 (0.5)	6.3 (0.5)	6.3 (0.5)

(b) *RMSE*

Table 4

Errors for Circle-Arc Equating (Means and Standard Errors)

V	Mini	Semi	Midi
$N = 10$			
20	0.0 (3.7)	0.0 (3.7)	−0.2 (3.7)
25	0.0 (3.2)	−0.2 (3.3)	0.0 (3.3)
30	−0.1 (3.1)	−0.2 (3.3)	−0.3 (3.1)
35	0.0 (2.9)	−0.3 (3.0)	−0.3 (2.9)
40	0.0 (2.8)	−0.1 (2.7)	−0.2 (2.9)
$N = 25$			
20	−0.1 (2.4)	0.0 (2.5)	−0.3 (2.5)
25	−0.1 (2.2)	−0.2 (2.2)	−0.2 (2.2)
30	0.0 (2.1)	−0.1 (2.0)	−0.2 (2.1)
35	0.0 (1.9)	−0.1 (1.9)	−0.2 (1.9)
40	0.0 (1.8)	0.0 (1.8)	−0.2 (1.8)
$N = 50$			
20	0.0 (1.8)	−0.1 (1.9)	−0.1 (1.8)
25	0.1 (1.6)	0.0 (1.7)	−0.1 (1.6)
30	0.0 (1.5)	0.0 (1.5)	−0.1 (1.5)
35	0.0 (1.5)	−0.1 (1.4)	−0.1 (1.5)
40	0.1 (1.4)	0.0 (1.4)	−0.1 (1.4)
$N = 100$			
20	0.0 (1.4)	−0.1 (1.3)	−0.1 (1.4)
25	0.0 (1.2)	−0.1 (1.3)	0.0 (1.3)
30	0.0 (1.1)	−0.1 (1.1)	−0.1 (1.2)
35	0.0 (1.1)	0.0 (1.1)	−0.1 (1.1)
40	0.0 (1.1)	−0.1 (1.0)	−0.1 (1.0)

(a) *Bias*

V	Mini	Semi	Midi
$N = 10$			
20	6.8 (1.8)	6.8 (1.7)	6.9 (1.8)
25	6.7 (1.6)	6.7 (1.6)	6.7 (1.6)
30	6.5 (1.6)	6.7 (1.6)	6.6 (1.5)
35	6.5 (1.5)	6.5 (1.6)	6.6 (1.5)
40	6.4 (1.5)	6.5 (1.5)	6.5 (1.5)
$N = 25$			
20	6.6 (1.0)	6.6 (1.1)	6.6 (1.1)
25	6.5 (1.0)	6.5 (1.0)	6.5 (1.1)
30	6.5 (1.0)	6.5 (1.0)	6.5 (1.0)
35	6.4 (1.0)	6.4 (1.0)	6.4 (0.9)
40	6.4 (0.9)	6.4 (0.9)	6.3 (1.0)
$N = 50$			
20	6.5 (0.7)	6.5 (0.7)	6.5 (0.7)
25	6.4 (0.7)	6.4 (0.7)	6.4 (0.7)
30	6.4 (0.7)	6.4 (0.7)	6.4 (0.7)
35	6.4 (0.7)	6.4 (0.7)	6.4 (0.7)
40	6.3 (0.7)	6.4 (0.7)	6.3 (0.7)
$N = 100$			
20	6.4 (0.5)	6.4 (0.5)	6.4 (0.5)
25	6.4 (0.5)	6.4 (0.5)	6.4 (0.5)
30	6.4 (0.5)	6.3 (0.5)	6.4 (0.5)
35	6.3 (0.5)	6.3 (0.5)	6.3 (0.5)
40	6.3 (0.5)	6.3 (0.5)	6.3 (0.5)

(b) *RMSE*

Figure 1

Distributions of Differences

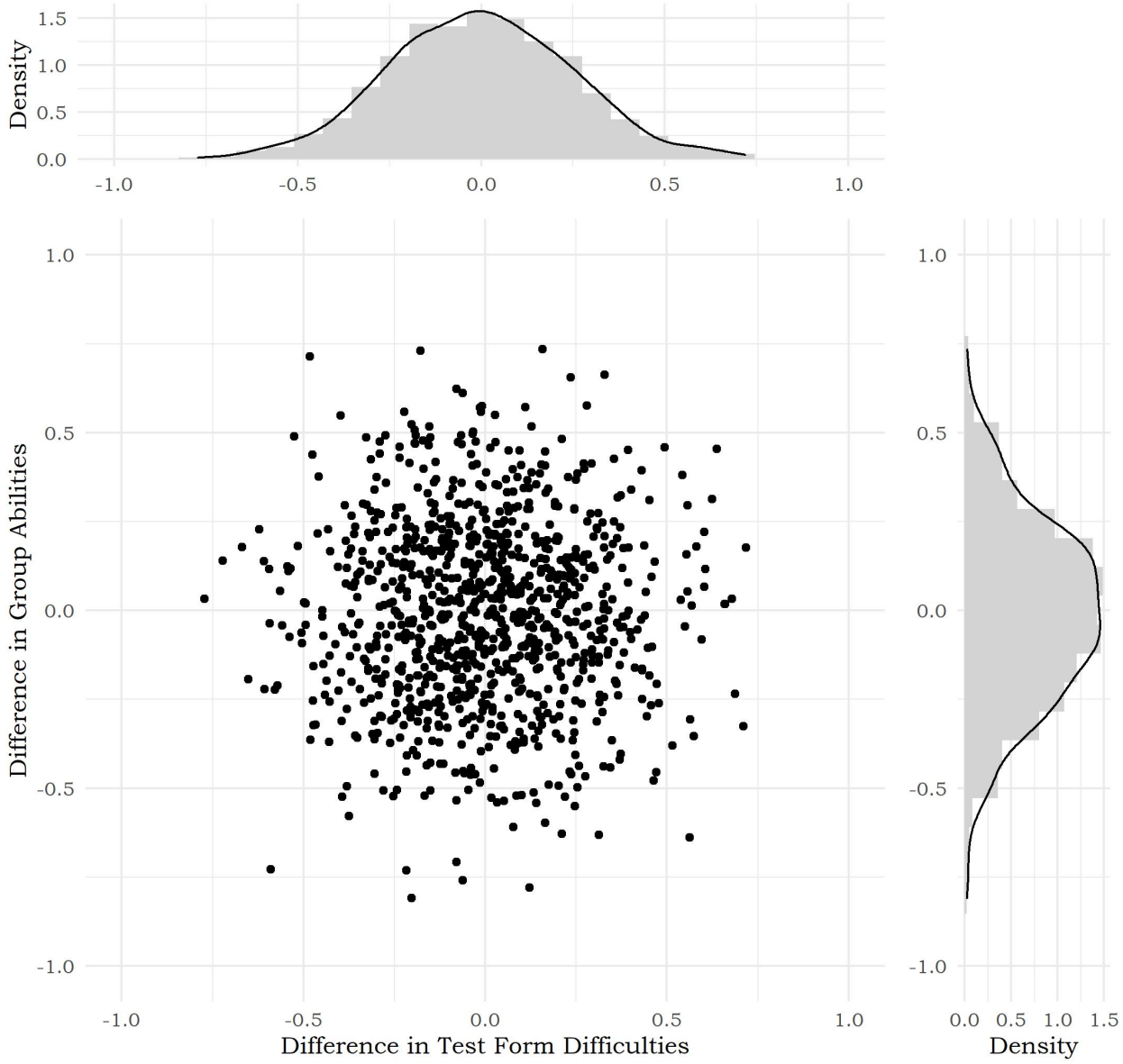

Figure 2

Bias by Anchor Length

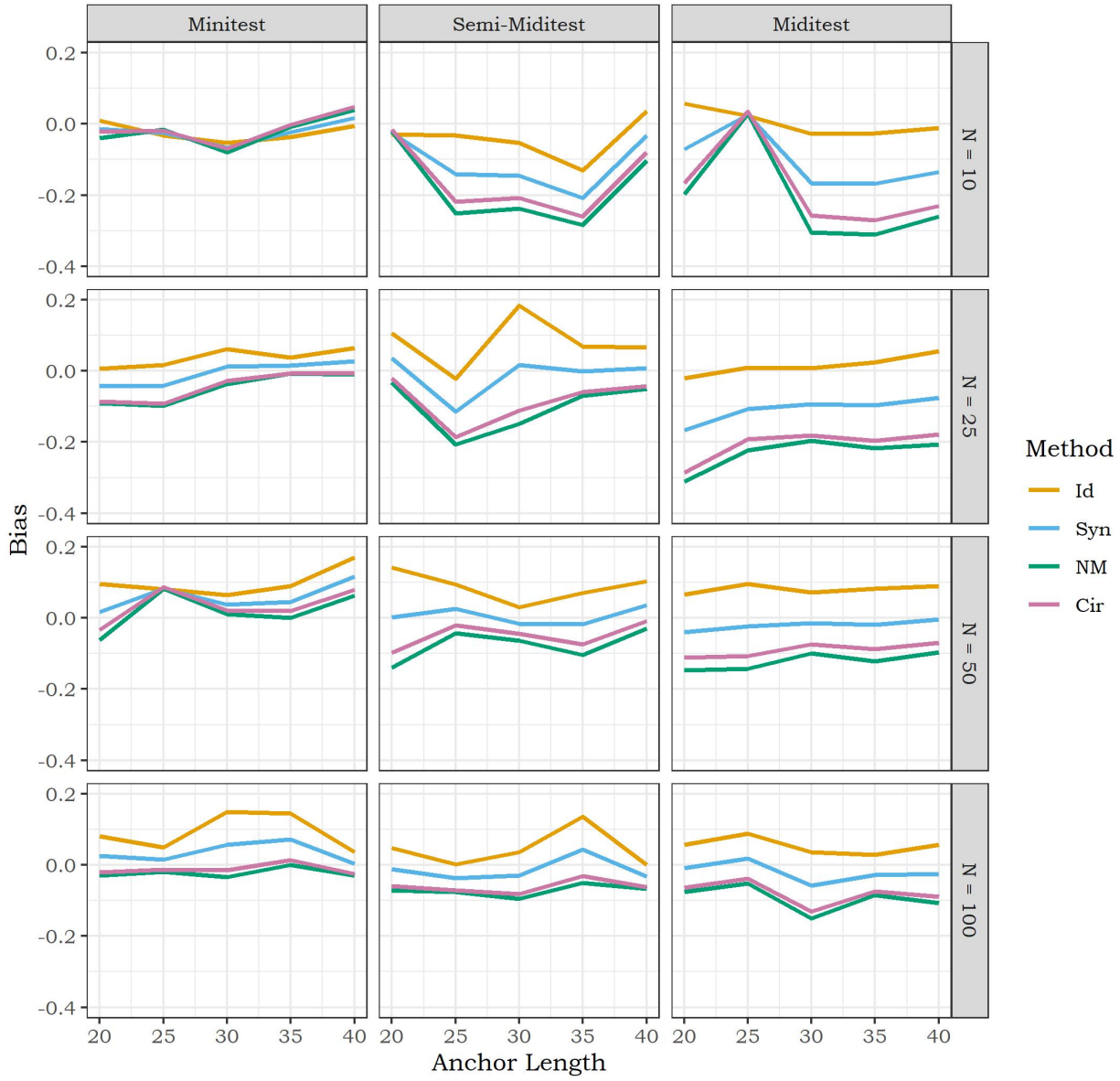

Figure 3

RMSE by Anchor Length