When to Use Synthetic Linking Functions 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

In small-sample equating, a synthetic linking function may be used to reduce random equating error. However, this benefit comes at a cost: an increase in bias when test forms differ in difficulty. This study aims to identify which, if any, situations are best handled by synthetic equating. Test forms were simulated to contain 100 total items, of which 30 items comprised the anchor. Four sample sizes were considered ($N = 10, 25, 50, 100$), and the difference in test form difficulties as well as the difference between group abilities were treated as random to encompass a variety of situations. Results indicate that synthetic equating may be preferable when the sample size is 25 or smaller and when it is known that the test forms do not differ markedly in difficulty. In all other cases, however, it is recommended that traditional equating functions be used.

Keywords: small-sample equating, synthetic equating

When to Use Synthetic Linking Functions in Small-Sample Equating

Whenever samples are used to estimate an equating relationship, random equating error is present. Random equating error is typically indexed by the standard error of equating (SEE), which may be thought of as the standard deviation of equated scores across hypothetical replications. Because the SEE is known to increase as sample size decreases, it becomes especially concerning in situations where samples are small (Kolen & Brennan, 2014; Parshall, Du Bose Houghton, & Kromrey, 1995; Skaggs, 2005).

One way to minimize random equating error is by using the identity function. This function assumes that the two test forms being equated are completely parallel. Thus, an observed score on the new form (X) is considered to be equivalent to the same observed score on the old form (Y). In other words, identity equating is analogous to not equating at all.

$$id_Y(x) = x \tag{1}$$

When the identity function is used, the SEE is zero. This is true because regardless of the samples that are used or the sample sizes, the estimated equating relationship remains unchanged. Consequently, there is no variation in the equated scores that are produced. Rather, the only error that is present is the systematic error, which represents the difference between the estimated equating relationship and the true equating relationship.

Due to the assumption that test forms are parallel, identity equating has been shown to perform well when samples are small and the test forms are similar in difficulty (Kim, von Davier, & Haberman, 2008). However, as the difference between test form difficulties increases, the results of identity equating become more and more biased. Eventually, the benefits of using identity equating become negated as the systematic error increases. In such cases, one might consider using a synthetic linking function, which computes a weighted average of the identity function and a second equating function (Kim

et al., 2008). This may be written as

$$syn_Y(x) = we_Y(x) + (1 - w)id_Y(x) \tag{2}$$

where w is a weight between 0 and 1, e is an equating function other than the identity function, and id is the identity function.

As a weighted average, the synthetic linking function attempts to balance its two components. The inclusion of the identity function is aimed at reducing the random equating error, while the inclusion of the equating function is aimed at reducing the systematic error. A natural question is how to compute the optimal weights such that the total error is minimized. A general rule of thumb is to increase the weight placed on the identity function when the test forms are constructed to be similar in difficulty, but to decrease this weight as the sample size increases. Yet, unless the two functions are parallel, it is unlikely that the chosen weights will maintain the symmetry property of equating. If the symmetry property is desired, then there is guaranteed to exist at least one set of weights that satisfy this requirement (Holland & Strawderman, 2011). In particular, if the equating function is linear, then these weights may be computed as

$$w = \frac{(1 + b_e^2)^{-1/2}}{(1 + b_e^2)^{-1/2} + (1 + b_{id}^2)^{-1/2}} \tag{3}$$

where b_e is the slope of the equating function, and b_{id} is the slope of the identity function.

Though the synthetic linking function offers a practical compromise between two alternatives, its true utility lies in whether it is able to outperform both individual functions. Kim et al. (2008) conducted two real-data studies in which the test forms were designed to be nearly parallel. They compared the results of identity equating, chained linear equating, and synthetic linking for five total sample sizes ($N = 10, 25, 50, 100, 200$). When conducting the synthetic linking, they assigned equal weights ($w = 0.5$) to the chained linear equating function and to the identity function. Generally, they found that synthetic linking produced bias and root mean squared errors that were between those observed for identity equating and those observed for chained linear equating.

In an ideal situation, the two test forms being equated would be as close to parallel as possible. However, this is difficult to achieve in practice. To address a different type of scenario, Kim, von Davier, and Haberman (2011) considered situations in which the two test forms were clearly not parallel. They used real data and compared the results of identity equating, mean equating, Tucker equating, Levine-observed score equating, chained linear equating, and synthetic linking for sample sizes ranging from 19 to 70 examinees. For synthetic linking, two different weight systems were employed: (1) equal weights ($w = 0.5$) and (2) placing a heavier weight on the equating function ($w = 0.7$) than on the identity function. As expected, identity equating suffered across all conditions since the test forms differed markedly in difficulty. Consequently, the synthetic linking function that placed a heavier weight on the equating function performed better than the synthetic linking function that used equal weights. In fact, for samples having fewer than 25 examinees, the synthetic linking function with unequal weights produced the lowest root mean squared differences of all the considered methods.

Though Kim et al. (2008) and Kim et al. (2011) were able to provide reasonable estimates regarding the differences in their test form difficulties, the true differences remained unknown since the studies used real data. One remedy for this uncertainty is to conduct a simulation study in which the true difference between test form difficulties is known. Babcock, Albano, and Raymond (2012) conducted such a study in which they manipulated equating method, sample size ($N = 20, 50, 80$), the difference in test form difficulties, and the difference in group abilities. For synthetic linking, they used the Tucker linear equating function, which was given the heavier weight ($w = 0.75$). They found that when the test forms were equal in difficulty, identity equating performed better than all other methods, regardless of sample size or the difference in group abilities. When the test forms differed in difficulty, mean equating or nominal weights mean equating tended to be the most effective methods. As a result, there were no situations in which they found synthetic linking to be the most effective method.

Similar to previous studies, the purpose of this study is to identify which, if any, small-sample situations are best handled by synthetic equating. However, whereas prior research has focused on real data or has treated the difference in test form difficulties as a fixed effect, I treat the difference in test form difficulties as a random effect, thereby allowing for a wide range of plausible scenarios. Additionally, because prior research has focused on examining a limited number of conditions, the authors had been able to think carefully before selecting the weights that would be used inside the synthetic linking function. However, because this study considers a large number of scenarios, I decided to use weights that satisfy the symmetry property of equating rather than separately considering each case. Thus, I refer to this process as *synthetic equating* rather than using the more general term, synthetic linking.

Method

Data

All 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 represent the data of several testing programs (e.g., Babcock et al., 2012). Both the old and new test forms were simulated to contain 100 items. Of these, 30 items comprised the anchor and were common between the two forms. This anchor length was selected as it follows the recommendation given by Kolen and Brennan (2014) for a test of this size. Four sample sizes were considered in total ($N = 10, 25, 50, 100$), and the difference between test form difficulties as well as the difference between group abilities were treated as random to encompass a variety of situations. Additionally, 1,000 replications were conducted for each sample size.

For both 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 followed $N(\Delta_b, 1)$ where Δ_b

represents the difference in test form difficulties. Specifically, Δ_b was sampled from $N(0, 0.25)$ to allow for instances where the new form was more difficult than the old form, and vice versa. Similarly, the ability parameters of the old group followed $N(0, 1)$, while the ability parameters of the new group followed $N(\Delta_\theta, 1)$ where Δ_θ represents the difference in group abilities. More specifically, Δ_θ was sampled from $N(0, 0.25)$.

After simulating the old form data, the 30 anchor items were randomly selected such that the following criteria were satisfied: (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 the standard deviation of the old form p-values. This selection process follows the recommendation that an anchor test be representative of the total test in terms of its statistical characteristics (Kolen & Brennan, 2014). Once the anchor items had been selected, the new form data were simulated such that the entire test form approximated the distribution described above.

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 two methods, Tucker mean equating and Tucker linear equating, assume (1) the regressions of the total test scores on the anchor test scores use the same linear functions in the old and new populations, and (2) the conditional variances of the total test scores given the anchor test scores are equal in both populations, as well (Gulliksen, 1950). One drawback of these methods is that they require the estimation of variance and covariance terms that may not be accurate when sample sizes are small. Hence, nominal weights mean equating simplifies the computation by omitting these values and replacing them with a ratio of the number of

total test items to the number of anchor test items (Babcock et al., 2012). The fourth method, chained linear equating, uses a different idea entirely. Chained linear equating places the new form onto the scale of the anchor test, then places the anchor test onto the scale of the old form (Angoff, 1971; Holland & Dorans, 2006). Two advantages of this method are that it is intuitive and it is relatively straightforward to compute. Moreover, all methods were selected because they have been shown to perform reasonably well in situations where samples are small (e.g., Babcock et al., 2012; Kim et al., 2011).

Each equating method was evaluated twice. First, it was paired with the identity function to represent synthetic equating, where weights were chosen such that the symmetry property of equating was maintained (see Equation 3). Second, each method was considered on its own (i.e., without the identity function) to represent traditional equating. Results were then compared across equating types (identity, synthetic, traditional) and equating methods (nominal weights mean, Tucker mean, Tucker linear, chained linear).

Evaluation Criteria

In order to compare the equated scores across replications, two criteria were assessed: bias and the root mean squared error (RMSE). As an index of systematic error, bias quantifies the difference between an estimated equating relationship and the true equating relationship. The bias of an equating function may be written as

$$\text{Bias} = \frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j - y_j) \tag{4}$$

where j denotes an examinee, N is the total number of examinees, $\hat{y}_j$ is an estimated equated score, and y_j is a true score. In contrast, the RMSE is an index of total error, as it incorporates both the random error of equating and systematic error.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j - y_j)^2} \tag{5}$$

In this study, an examinee's true score was defined as the one they received after being simulated to take the old form. Then for each replication and equating method, the

bias and RMSE were compared across sample sizes, differences in test form difficulties, and differences in group abilities. The methods that were more effective should display 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.

Figures 2 and 3 plot the bias and RMSE of each equating method against the difference in test form difficulties. Similarly, Figures 4 and 5 use the difference in group abilities. For consistency, all figures display the difference of interest along the x-axis, and the criterion of interest along the y-axis. Furthermore, the first column in each figure uses one line to represent each of the three equating types: identity (Id), synthetic (Syn), and traditional (Trad). For synthetic and traditional equating in particular, these lines represent the average across the four equating methods. In contrast, the second and third columns use one line to represent each of the four equating methods: nominal weights mean (NM), Tucker mean (TM), Tucker linear (TL), and chained linear (CL). Specifically, the second column presents the results for each of the synthetic equating methods, while the third column presents the results for each of the traditional equating methods.

Differences in Test Form Difficulties

Across all sample sizes, the identity and synthetic equating methods yielded large amounts of bias when the test forms differed even slightly in difficulty (Figure 2). In fact, when the absolute difference in test form difficulties was 0.25, identity equating tended to over- or underestimate an examinee's score by more than 2 points (out of 100 total points). Synthetic equating performed slightly better in this regard, though not by much. It tended

to over- or underestimate an examinee's score by approximately 1.5 points. The traditional equating methods displayed very small amounts of bias overall, regardless of whether the test forms were similar or had different difficulties. Moreover, these methods showed a slight decrease in bias as sample size increased.

With respect to the RMSE, it was generally the case that synthetic equating was preferred when the test forms were similar in difficulty and sample sizes were small (Figure 3). However, as sample size increased, the window in which synthetic equating was preferred quickly narrowed. Consider, for example, when the sample size was 10. Synthetic equating was generally preferred when the absolute difference in test form difficulties was less than 0.375. Else, traditional equating was preferred for the more extreme cases. However, when the sample size increased to 100, there were far more instances where traditional equating was the best option. Moreover, the benefits of using traditional equating became even more pronounced as the difference in test form difficulties increased.

For traditional equating in particular, each of the four methods performed similarly with respect to the bias. However, noticeable differences emerged when considering the RMSE, especially when sample sizes were small. Generally, Tucker mean and nominal weights mean equating yielded the lowest RMSEs. They performed quite similarly, as did Tucker linear and chained linear equating. In contrast, the bias and RMSE for the four synthetic equating methods were nearly indistinguishable, suggesting that none of the synthetic equating methods offered a clear advantage over the others.

Differences in Group Abilities

When groups were similar in ability, identity equating, synthetic equating, and traditional equating yielded similar amounts of bias (Figure 4). However, as differences in group abilities became more pronounced, the traditional equating methods tended to exhibit more bias than the other methods, on average. Moreover, because the bias of the synthetic equating methods fell between that of identity equating and that of the

traditional equating methods, the bias of the synthetic equating methods also suffered as group differences increased.

The RMSE results varied by sample size. When the sample size was 10, synthetic equating yielded the smallest RMSEs, regardless of the difference in group abilities (Figure 5). Then, when the sample size increased to 25, synthetic equating and traditional equating performed very similarly, suggesting that either would be a reasonable choice. However, for sample sizes of 50 and 100, the traditional equating methods almost always yielded the smallest RMSEs. This held true for all of the group differences that were considered. Another consistent result was that identity equating yielded the largest RMSEs of any equating type, regardless of group differences or sample size.

When considering specific equating methods, nominal weights mean equating tended to yield the smallest amounts of bias, especially for sample sizes that were 25 or larger. With respect to the RMSE, nominal weights mean equating and Tucker mean equating tended to produce the smallest values. These patterns were observed for both synthetic equating and traditional equating, though the differences between the traditional equating methods were generally more pronounced.

Discussion

As is the case with most statistical procedures, test equating often undergoes adjustments when samples are small. One adjustment that has been suggested is the use of a synthetic linking function as opposed to a traditional equating function. Because a synthetic linking function is a weighted average of the identity function and a second equating function, it is often described as a compromise between the two methods. In particular, the hope is that the individual functions are able to compensate for one another's weaknesses, thereby reducing the total error.

Prior research has examined the effectiveness of synthetic linking functions for specific tests and specific conditions. However, this study treated the difference in test form

difficulties as a random effect, thereby allowing for a wider range of plausible conditions. In addition, because so many different scenarios were considered, I used weights that satisfied the symmetry property of equating whereas previous studies did not. Thus, the purpose of this study was to consider the general effectiveness of synthetic equating when samples are small.

Results showed that synthetic equating yielded large amounts of bias and large RMSEs when test forms differed even slightly in difficulty. This held true across all sample sizes, and traditional equating methods consistently performed better with respect to both criteria. When the test forms were similar in difficulty, identity equating, synthetic equating, and traditional equating all yielded very small amounts of bias. The synthetic equating methods tended to produce smaller RMSEs than the traditional equating methods, though both equating types performed similarly for sample sizes of 100.

The difference in group abilities did not seem to have as large of an effect. The bias of all three equating types were generally comparable, though identity equating offered a slight advantage when the difference in group abilities was large. The synthetic equating methods yielded some of the smallest RMSEs when the sample size was 10 or 25, regardless of group differences. However, when the sample size was 50 or 100, the traditional equating methods tended to be more accurate.

In general, it seems that synthetic equating should only be considered when the sample size is 25 or smaller and when it is known that the test forms do not differ markedly in difficulty. In all other cases, however, it may be wise to use traditional equating functions, as the benefits of synthetic equating are unlikely to outweigh the costs. When selecting an equating function, nominal weights mean equating and Tucker mean equating tended to be the most effective, especially when samples were very small. However, as sample size increased, the differences between equating methods generally became less noticeable.

This study has several limitations that could affect the generalizability of its results. First, both test forms were simulated to contain 100 total items, where 30 items comprised the anchor. Tests having different lengths or compositions would likely produce different results. Second, the anchor items were selected in a way such that certain restrictions were followed. If these restrictions were altered, then the results could differ, as well. Third, I considered the case in which both groups contained the same number of examinees. In practice, it is likely that one group would be larger than the other, making this assumption somewhat unrealistic. Fourth, the weights used within the synthetic linking function were chosen such that the symmetry property of equating was maintained. While such weights offer certain advantages, testing programs may wish to consider different weights depending on their particular needs. Fifth, I considered four equating methods that are often cited in the small-sample equating literature. These methods are by no means exhaustive, and as new methods are developed, it would be useful to include them, as well. Future research could address any one of these limitations, specifically by considering different test lengths, anchor item selection methods, sample sizes, weights, and equating methods.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

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.

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., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger Publishers.

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., 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

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement*, *32*(1), 37–54.

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

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

Figure 1

Distributions of Differences

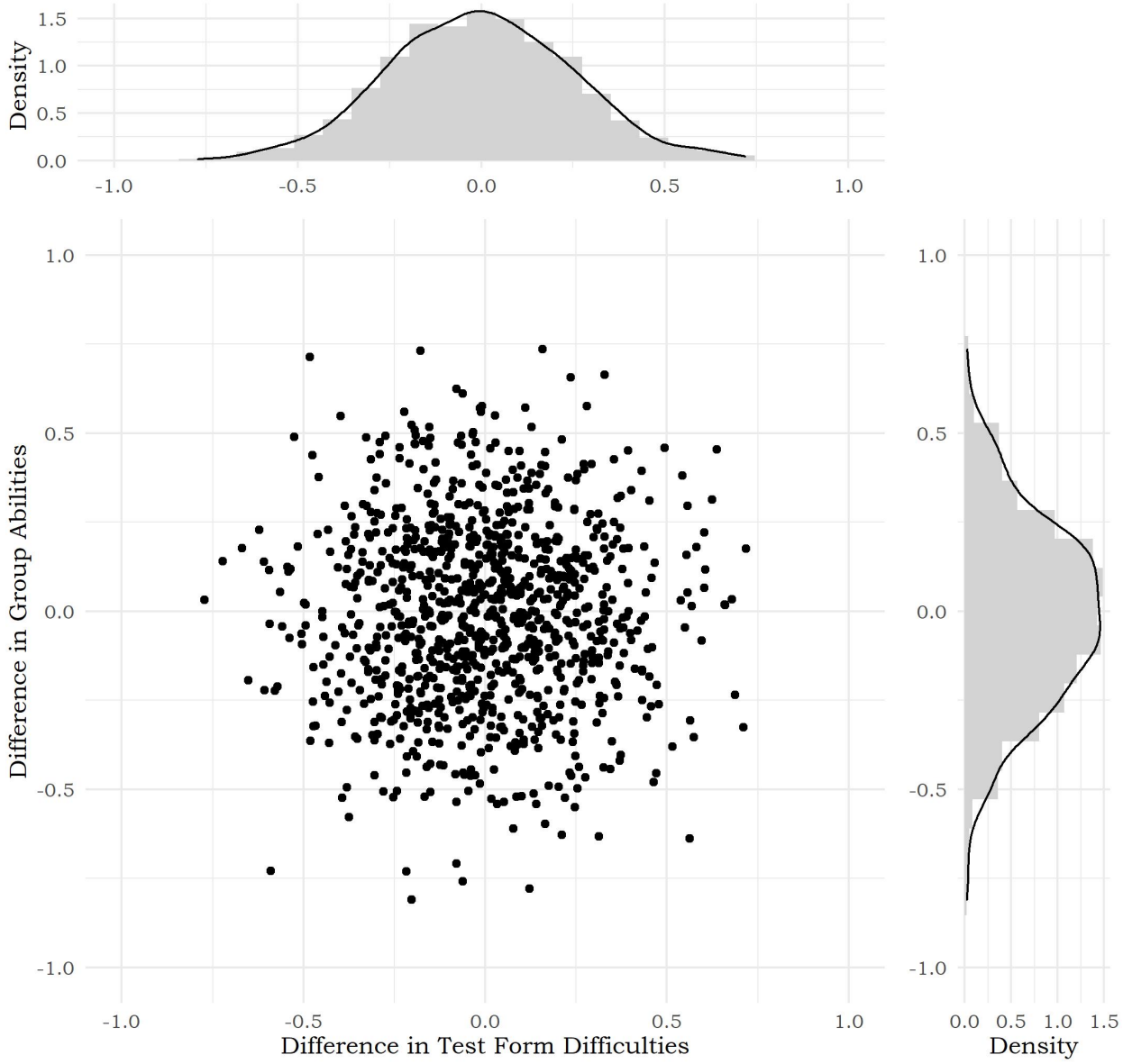

Figure 2

Bias by Difference in Test Form Difficulties

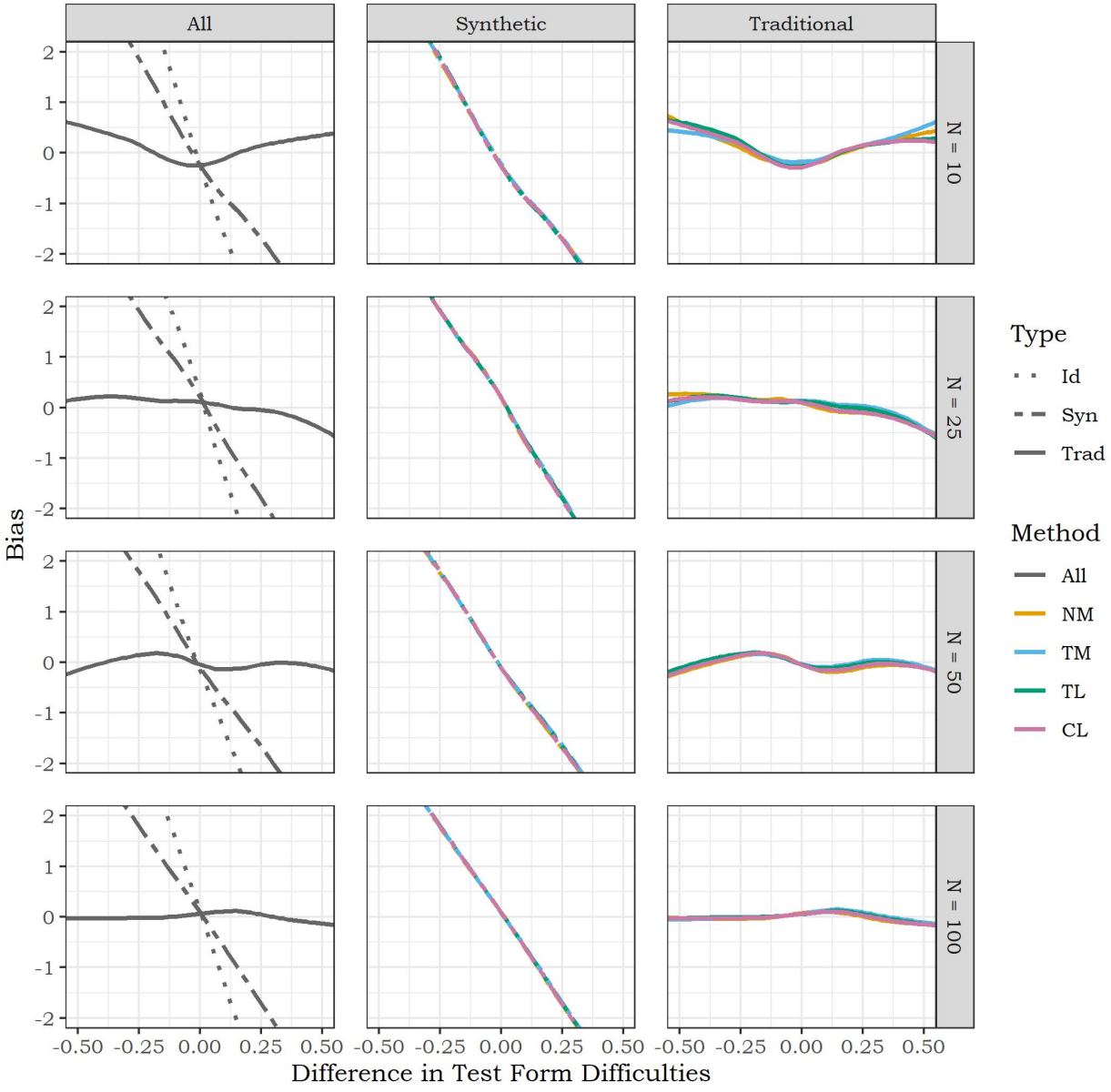

Figure 3

RMSE by Difference in Test Form Difficulties

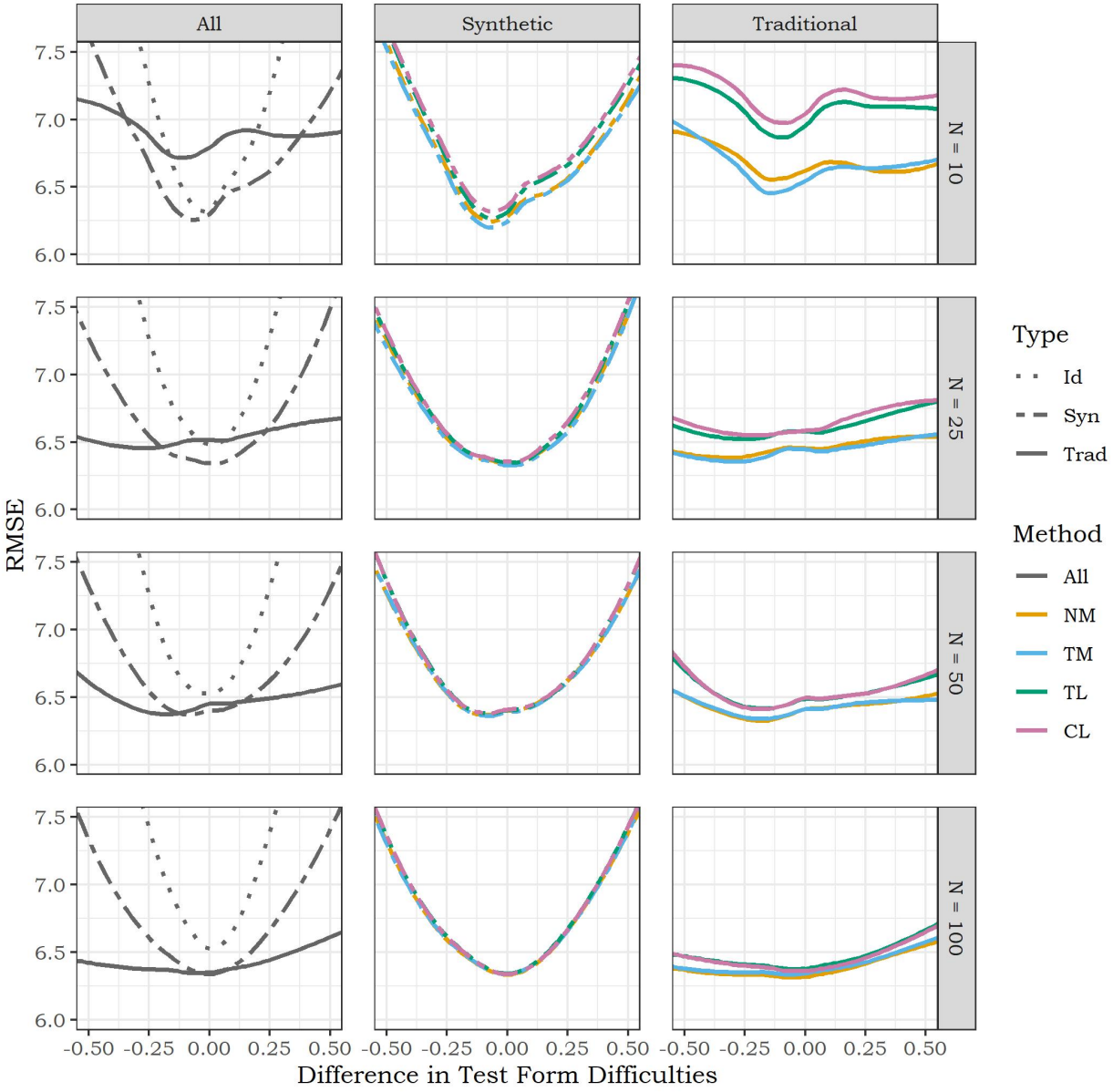

Figure 4

Bias by Difference in Group Abilities

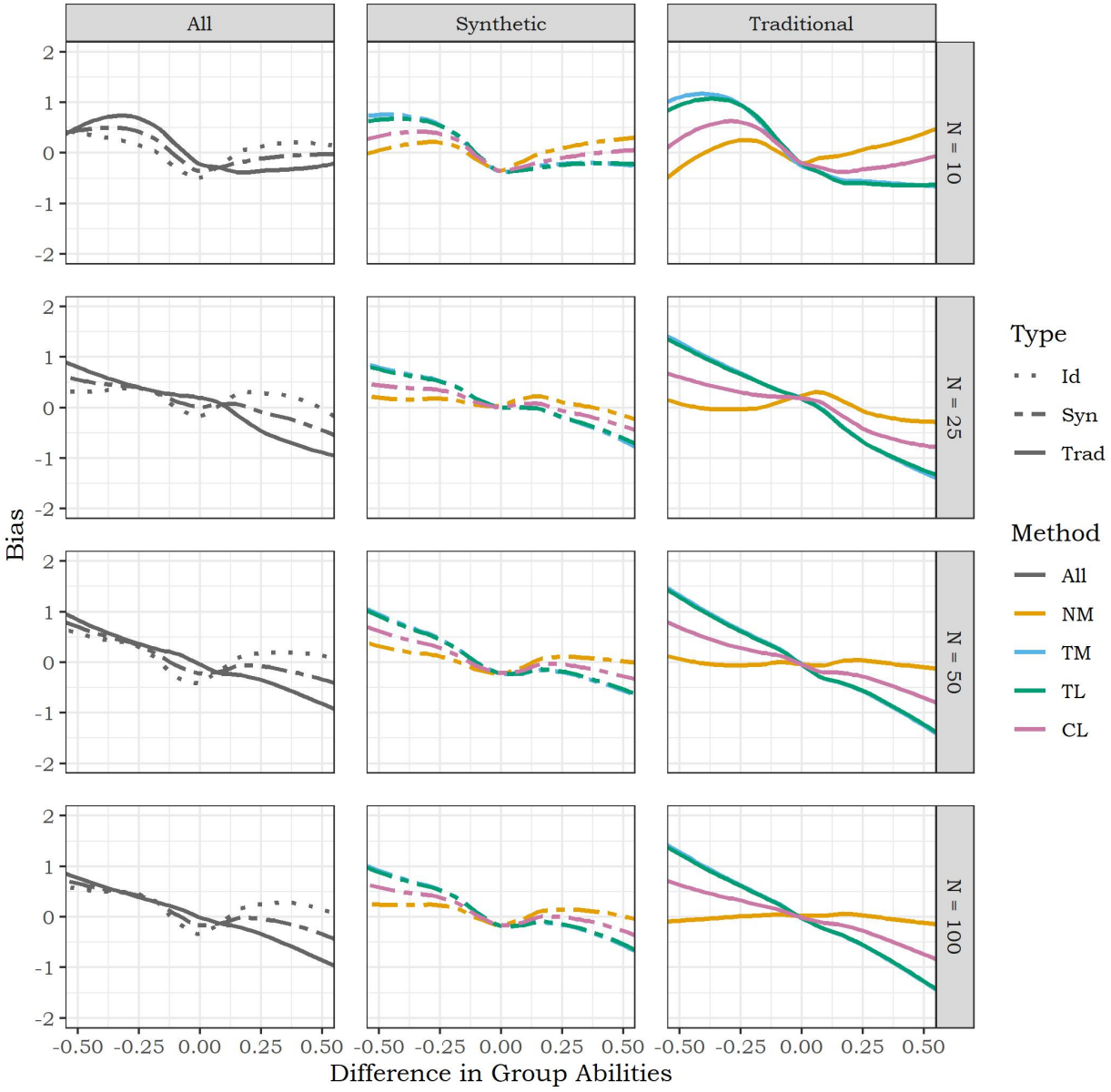

Figure 5

RMSE by Difference in Group Abilities