Research Report

Effect of Statistically Matching Equating Samples for Common-Item Equating

ETS RR–21-02

Ru Lu
Sooyeon Kim

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Effect of Statistically Matching Equating Samples for Common-Item Equating

Ru Lu & Sooyeon Kim

Educational Testing Service, Princeton, NJ

This study evaluated the impact of subgroup weighting for equating through a common-item anchor. We used data from a single test form to create two research forms for which the equating relationship was known. The results showed that equating was most accurate when the new form and reference form samples were weighted to be similar to the target population. When the target population was a combination of the two equating samples and one sample was weighted to be similar to the other, the equating was less accurate but still much more accurate than equating with unweighted samples.

Keywords Equating; target population; subgroup; weighting

doi:10.1002/ets2.12313

For most large-scale international examinations, such as the *GRE*®, *TOEFL*®, and *TOEIC*® tests, the target testing population tends to be composed of multiple subpopulations based on test takers' characteristics, such as testing experience, geographic region, and native languages. When subgroups of test takers have different levels of the skills measured by the test, the equating function can vary depending on the subgroup compositions (Kolen & Brennan, 2004, p. 285). The common dilemma associated with the choice of equating samples is that including test takers from all the subgroups may reduce the overall accuracy and stability of equating, whereas removing a particular subgroup from the equating sample may reduce sample representativeness.

Several studies showed that the choice of equating samples could impact the resulting equating function. Liang et al. (2009) discussed the impact of including international test takers when equating a domestically used test. Using data from the *PSAT/NMSQT*® administrations, they examined what happens if the proportion of the group of English as a second language (ESL) examinees increases in the population. They created synthetic samples from the available data by varying percentages of ESL examinees and ran equating analyses on these synthetic samples to assess the sensitivity of the equating results to the different ESL composition in the samples. They compared the total group equating to the ESL subgroup equating using a subpopulation invariance perspective. Although the difference was small on average, not all conditional differences between the total group conversion and subgroup conversions were negligible, indicating the impact of subgroup composition on the equating. Kim and Walker (2009) investigated the population invariance of equating by comparing equating functions derived using two distinct subgroups (first-time test taker versus repeater) from the examinee population for a large-scale licensure test. In their comparisons, equating functions remained essentially invariant across all repeaters versus first-time test takers, supporting score equitability of the two forms. However, when the repeater subgroup was subdivided based on the particular form examinees took previously, subgroup equating functions substantially differed from the total-group equating function, indicating subgroup dependency of score equating.

For testing programs that have multiple administrations across years, the common-item design (often called nonequivalent groups with anchor test [NEAT]) is one of the most popular data collection designs to conduct equating. With the common-item design, psychometricians face challenges when the new and reference form samples differ in their ability levels. The issue becomes more complicated when the proportions of distinct subgroups in the equating samples vary over administrations. The reason for the variation in subgroup composition of the equating sample might be either a permanent condition (e.g., population shift) or unexpected events (e.g., students could not take the test as scheduled). When subgroups differ in their performance levels and their proportions vary over administrations, testing programs

Corresponding author: R. Lu, E-mail: rlu@ets.org

often use the data from a particular (major) subgroup to create the conversions from which all test takers' reported scores are derived (e.g., Duong & von Davier, 2012; Liang et al., 2009). This practice has a psychometric benefit in that using the data from a particular (major) subgroup may make the item calibration and linking more stable and secure across administrations than using the data from the entire group (Kim & Robin, 2017). Despite its psychometric benefit, however, failure to include nonmajor subgroups in the operational analysis may bias the linking as an estimate of the linking in the full population. Any potential bias associated with the use of a particular subgroup conversion would increase if the proportion of the excluded subgroups becomes larger over time or if the proportion of subgroups is inconsistent over administrations.

Another way to address this problem is to include all test takers, but with weights, so that the two groups of equating samples would be equivalent in terms of subgroup composition. For example, in the context of the equivalent groups design, Duong and von Davier (2012) statistically compared the equating functions under different weighting and verified that the application of the right weights (e.g., inversely proportional to the sample sizes) to the two distinct subgroups becomes more advantageous in situations where sample sizes are not equal and group ability differences are large. Livingston (2014) tried two procedures for using demographic information to achieve groups equivalent in ability, referred to as "demographically adjusted groups" (DAG). In one procedure, the composite variable that best predicted test scores was used to obtain student weights. In the other procedure propensity scores were used to determine the weights. He found that although both procedures greatly reduced the ability difference between the groups, neither procedure was adequate for creating equivalent groups.

Instead of using selected demographic information to create equivalent groups, Haberman (2015) used a large set of background information on individual test takers with the adjustment by minimum discriminant information method to create "pseudo-equivalent groups" (PEG) and then conducted linking with the equivalent groups design. With an established testing program that uses common items for equating, he verified that PEG linking produces results similar to the conventional equatings with anchor tests. As an extension of the PEG modeling, Lu and Guo (2018) further proposed to use both background information and anchor test scores to adjust group ability difference. With simulation studies based on real data, they showed that the equating based on common items could be improved by incorporating the PEG adjustment procedure into the NEAT process. With common item designs, Qian et al. (2013) studied the effects of weighting test takers using item response theory (IRT) true-score equating. They found that the scale transformation parameters for converting the item parameter estimates from one scale to another were more stable with weighted equating than with unweighted equating.

Purpose of the Study

In practice, not all testing programs collect a large set of test takers' background information. For testing programs that normally rely on common items for equating, special circumstances might cause a one-time increase or decrease in the number of test takers in a particular subgroup. This study attempts to address such an equating situation, especially when the subgroups differ in the ability assessed by the test. This study tries to answer the following research questions:

1. Does weighting the equating samples lead to more accurate equating than using unweighted samples?
2. Do different weighting schemes lead to different equating results? If so, which weighting scheme outperforms in terms of equating accuracy? For this question, we compare three different weighting schemes.

To answer these two questions, we created a hypothetical equating situation, using real test data, in which the population of test takers consisted of two distinct subgroups that were represented in different proportions at different test administrations. We expect that when individual test takers in both new and reference form samples are weighted as a function of subgroup membership which is related to the performance on the test, the resulting equating would be more accurate.

Weighted Equating Using the Common Item Design

Assume that the testing population has two subpopulations P and Q. Samples P_x and Q_x took the new form X, and samples P_y and Q_y took the reference form Y. The data collection design for common-item equating with two subpopulations is presented in Table 1. Samples P_x and P_y come from subpopulation P; while Q_x and Q_y come from subpopulation Q. If

Table 1 Data Collection Design With Subgroups in Common Item Equating

Sample	New form X	Anchor test	Reference form Y
P_x	✓	✓	
Q_x	✓	✓	
P_y		✓	✓
Q_y		✓	✓

Note. ✓ denotes the presence of data.

the combined sample taking each form is representative of the target population, any of the conventional common-item equating methods, such as poststratification equating or chained equipercentile equating (Kolen & Brennan, 2004) would produce reasonable equating results.

With weighted equating, weights are applied to individuals in the equating sample to make the samples similar to the target population in terms of the proportion of subgroups. Each subgroup in each equating sample has a weight, and that weight is applied to each individual in that subgroup. Let w_{px} and w_{qx} represent the weights assigned to the samples P and Q taking the new form X, and w_{py} and w_{qy} represent the weights assigned to the samples P and Q taking the reference form Y. Let n_{px} and n_{qx} represent the sample sizes for the subgroups taking the new form X; and n_{py} and n_{qy} represent the sample sizes for the subgroups taking the reference form Y. Thus, for the new and reference form samples X and Y,

$$\frac{w_{px}n_{px}}{w_{qx}n_{qx}} = \frac{w_{py}n_{py}}{w_{qy}n_{qy}} = \frac{n_{pt}}{n_{qt}}. \quad (1)$$

Here $\frac{n_{pt}}{n_{qt}}$ indicates the subgroup ratio in the target population t. Solving the equation above with known sample sizes of each subgroup in the new and reference form samples and the subgroup ratio in the target population, we can obtain the weights.

After applying the appropriate weight to each individual in each subgroup of the equating samples, the score distributions of new and reference form samples are calculated. Then the conventional equating such as poststratification equating or chained equipercentile equating can be carried out using the weighted score distributions.

The Design of the Study

Data and Test Forms

We used data from a large-scale language testing program. The selected test form has 100 items and had been taken by about 140,000 test takers. Using the selected operational form as an item pool, we created a pair of research forms, similar in content but unequal in difficulty, with 20 items in common. Each form had 60 items. One form served as the new form and the other served as the reference form. By design, the new form was harder than the reference form, and their standardized mean difference (SMD) was about 0.25.

With operational data, we divided test takers into two subgroups based on their testing regions (Region 1 and Region 2). Table 2 presents the summary statistics for each region on the two research forms and the anchor test. Table 2 shows the difference between the two regions to be about 0.88 *SD* for Form X, 0.85 *SD* for Form Y, and 0.79 *SD* for the anchor. It also shows that the Region 1 test takers varied more in ability; the standard deviations for Region 1 are 17 – 22% larger than those for Region 2. The anchor-total correlations on the new and reference forms were .91 and .92, respectively, for Region 1; They were .90 and .92, respectively, for Region 2. In the following analysis, we treated the whole data set as the population for a resampling study with 100 replications.

Equating Condition

For the study, we created a special equating condition where the subgroup proportions changed dramatically from the reference form administration to the new form administration. Table 3 lists the number of test takers from each region in the new form sample, the reference form sample, and the combined sample. In the operational setting, the proportions of test takers from Regions 1 and 2 were about 23% and 77%, respectively. We specified those proportions as the reference

Table 2 Means and Standard Deviations of the Two Research Forms and the Anchor Test

Form/Test	Region 1 (n = 33,067)	Region 2 (n = 113,787)	All test takers (N = 146,854)
Form X (60 items)	37.5 (9.8)	45.5 (8.4)	43.7 (9.3)
Form Y (60 items)	40.1 (9.8)	47.7 (8.1)	46.0 (9.1)
Anchor (20 items)	12.7 (3.9)	15.5 (3.2)	14.9 (3.6)

Note. The values in parentheses indicate standard deviations.

Table 3 Subgroup Sample Sizes and Percentages in the Reference Form, New Form, and Combined Samples

Sample	Region 1	Region 2	Total
Reference form	5,000 (23%)	17,206 (77%)	22,206 (100%)
New form	10,000 (50%)	10,000 (50%)	20,000 (100%)
Combined	15,000 (36%)	27,206 (64%)	42,206 (100%)

Note. Combined sample = reference form sample plus new form sample.

form sample's subgroup proportions to make the reference form sample similar to the group of test takers that provided data for the study.

In each replication of the X-to-Y equating, we randomly selected, with replacement, 5,000 test takers from Region 1 and 17,206 test takers from Region 2 to construct the reference form sample. For the new form sample, we randomly selected, with replacement, 10,000 test takers from each region. The combined sample in the last row of Table 3 includes all test takers in the reference and new form, in which the proportions of test takers from Regions 1 and 2 in the combined sample were 36% and 64%, respectively.

Target Population

In this study, the purpose of weighting is to make the weighted samples similar to the target population in subgroup composition. For this purpose, we considered three different target populations:

1. A population similar to the reference form sample and to the operational data, with 23% of the test takers from Region 1 and 77% of the test takers from Region 2
2. A population similar to the new form sample, with 50% of the test takers from each region
3. A population similar to the combined sample, with 36% of the test takers from Region 1 and the remaining 64% from Region 2

Weighting Schemes

In each of the three target population conditions above, we carried out three weighted equatings and one unweighted equating. The three weighted equatings are as follows:

1. Weight the reference form sample to match the new form sample
2. Weight the new form sample to match the reference form sample
3. Weight both the new and reference form samples to match the combined sample

Thus, we investigated 12 conditions formed as a function of three target populations and four different equatings (three weighted and one unweighted).

Criterion Equating Functions

We created the two research forms from a single operational form. This form assembly enabled us to use single group direct equipercentile equating to obtain a criterion equating function. Recall that this study considers three different target populations whose subgroup compositions differ substantially. Because the relationship between the two research forms could differ across the three target populations, we created a separate criterion equating function for each target

Table 4 Three Target Population Conditions and Number of Test Takers Used to Obtain Criterion Functions

Target population	Region 1	Region 2	Total
Similar to reference form sample	33,067 (23%)	113,787 (77%)	146,854 (100%)
Similar to new form sample	33,067 (50%)	33,067 (50%)	66,134 (100%)
Similar to combined sample	33,067 (36%)	58,786 (64%)	91,853 (100%)

Note. The last column indicates the number of test takers who were used to obtain the criterion function in each target population condition.

population. We did this by choosing criterion equating samples from Regions 1 and 2 so that the proportions of test takers from the two regions matched the proportions in the hypothetical target populations.

Table 4 presents three target population conditions, their subgroup compositions, and the total numbers of test takers used to obtain the criterion equating functions. As shown, all test takers from Region 1 appeared in all three criterion equating samples. For each target population condition, we randomly selected test takers, without replacement, from Region 2 until the proportion of Region 2 test takers in the criterion sample matched the proportion in the respective target population. Then we created the criterion equating function by linking the two research forms using direct equipercentile equating in the group of combined test takers from Regions 1 and 2.

In addition, we computed two subgroup equating functions. Each of these subgroup equatings was a direct equipercentile equating in the group of all test takers from one of the two regions (33,067 for Region 1 and 113,787 for Region 2). We compared these two subgroup equatings to the three criterion equatings.

Evaluation Indices

In this study, common-item equating was replicated 100 times in each of the 12 equating conditions. All X-to-Y common-item equating functions (weighted and unweighted) were compared to their corresponding criterion X-to-Y function. In each condition, the differences were evaluated graphically and also were summarized as weighted averages across all raw score points. The three evaluation indices were average bias, standard error of equating (SEE), and root mean squared error (RMSE). These indices were calculated separately for each raw score point and across the entire score range. The following equations represent the bias, SEE, and RMSE measures conditioned on each raw score point, in which i indexes the raw-score values; j indexes the replications; $e_{yt}(x_i)$ represents the equipercentile equating function in the population t that is used to convert scores on new form X to the scale of the reference form Y; and $e_{yj}(x_i)$ is the estimated equating function based on common items of the jth replication.

$$Bias_i = \frac{\sum_{j=1}^{100} \left[e_{yj}\left(x_i\right) - e_{yt}\left(x_i\right) \right]}{100} \tag{2}$$

$$SEE_i = \sqrt{Var\left(e_y\left(x_i\right)\right)} \tag{3}$$

$$RMSE_i = \sqrt{Bias_i^2 + SEE_i^2} \tag{4}$$

As overall summary measures, we calculated the root mean squared bias, $\sqrt{\sum_i f_i Bias_i^2}$, SEE, $\sqrt{\sum_i f_i SEE_i^2}$, and RMSE, $\sqrt{\sum_i f_i RMSE_i^2}$, where f_i is the average frequency count at raw score i in the new form sample over 100 replications.

Results

All equatings were conducted with both poststratification equating and chained equating methods. Because poststratification equating and chained equating produced similar results, we present the results from the chained equating in the following sections for simplicity. We can offer the poststratification equating results upon request.

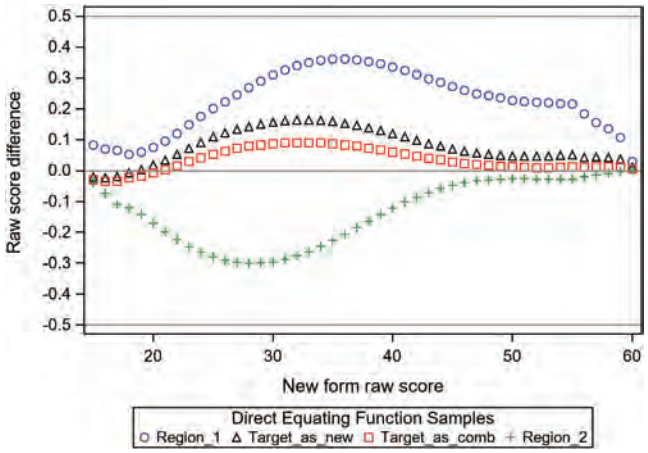

Figure 1 Criterion equating functions and subgroup equating functions.

Criterion Equating Functions and Subgroup Equating Functions

Among the three criterion equating and two subgroup-only equating functions, only the criterion equating function (Target = Reference Form) used the full operational data set for the equating functions. We used it as the base and compared the other four direct equatings with this one. Figure 1 presents the conditional equated raw score differences of each direct equipercentile equating using partial data from the full data direct equipercentile equating. In the figure, the circles represent the direct equipercentile equating using the Region 1 sample; the plus signs were the direct equating function using the Region 2 sample. The triangles and the squares represent the direct equipercentile equating when the new form sample and the combined sample were the target populations, respectively; they were closer to the base line than the two subgroup (i.e., region) only equatings. Figure 1 indicates that different compositions of subgroups in the target population produced different criterion equating functions. In the plot, the three criterion equating functions with different target populations do not differ by more than 0.2 raw-score points at any new-form raw score above 15.

Conditional Bias, SEE, and RMSE

Figure 2 presents the conditional bias plots for the three weighted equatings and the unweighted equating summarized over 100 replications. In each plot, the circles represent the weighted equating that adjusted the reference form sample to be similar to the new form sample; the triangles represent the weighted equating that adjusted the new form sample to be similar to the reference form sample; the squares represent the weighted equating that adjusted both the new and reference form samples to make them similar to the combined sample; and the plus signs represent the unweighted equating. In Figure 2(a), all equatings were compared with the criterion equating where the reference form sample was similar to the target population. The plot shows that the weighted equating that adjusted the new form sample to be similar to the reference form sample had almost zero bias for most of the raw score points. In Figure 2(b), all equatings were compared with the criterion equating where the new form sample was similar to the target population. It shows that the weighted equating that adjusted the reference form sample to be similar to the new form sample had almost zero bias for most of the raw score points. In Figure 2(c), all equatings were compared to the criterion equating where the combined sample was similar to the target population. Similarly, we observed that the weighted equating that adjusted both the new and reference form sample to be similar to the combined sample had almost zero bias. In all three plots, the unweighted equating had the largest bias for most of the raw score points. Plots in Figure 2 indicate that the weighting is doing what it is intended to do. When both equating samples are weighted to match the target population, there is almost no bias. When one sample is weighted to match the other, even if the other sample is not representative of the target population, the bias is much less than when equating without weighting either sample. Even the least appropriate of the three weighting methods reduces the bias to about one-fourth of the unweighted equating.

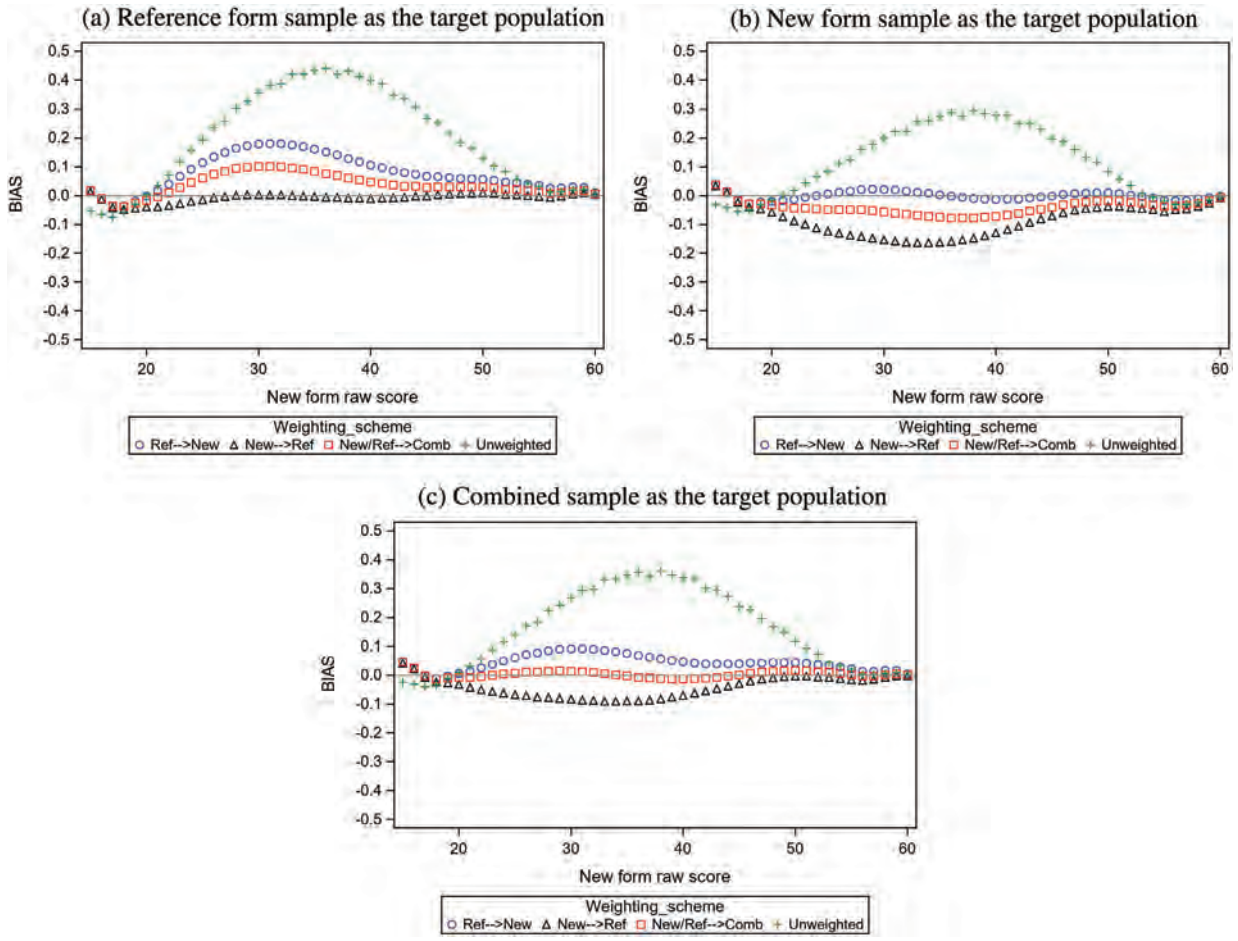

Figure 2 Conditional bias plots under three target population conditions.

Figure 3 presents the conditional SEE (CSEE) plots for the three weighted equatings and the unweighted equating summarized over 100 replications. It shows that the four equatings produced almost the same CSEE. Thus, weighting did not affect the SEE. This is probably because we had large sample sizes in our hypothetical equating condition. The conditional RMSEs are shown in Figure 4. Since the RMSE is a combination of the bias and the standard errors, the RMSE results closely parallel the results for bias.

Average Bias, SEE, and RMSE

Tables 5–7 summarize the three average deviance measures associated with all four weighted and unweighted equatings under each of the three target population conditions. As expected from the CSEE curves, the average SEEs (Table 6) of the equatings were also very similar in all conditions. Only the magnitudes of the bias and RMSE vary as a function of subgroup composition in the target population and the direction of weighting. As expected, the unweighted equatings that did not consider subgroup composition in the target population led to the largest bias and RMSE. In all three target population conditions, the average biases with unweighted equatings were about 0.2–0.3 raw score points. In contrast, among the three weighted equatings for each target population, the largest average bias was about 0.11 raw score points, and the smallest average bias was only 0.01 raw score points.

Figures 5 and 6 graphically present the overall bias and RMSE in each of the 12 conditions, respectively. They showed similar patterns as in the conditional biases and RMSEs. Thus, in general, among the three weighting schemes, the weighted equating that made both new and reference samples similar to the target population produced the least bias and RMSE. This trend was consistent with all three target population conditions. Even so, all three weighted equatings were more accurate than the unweighted equating for the research form used in this study.

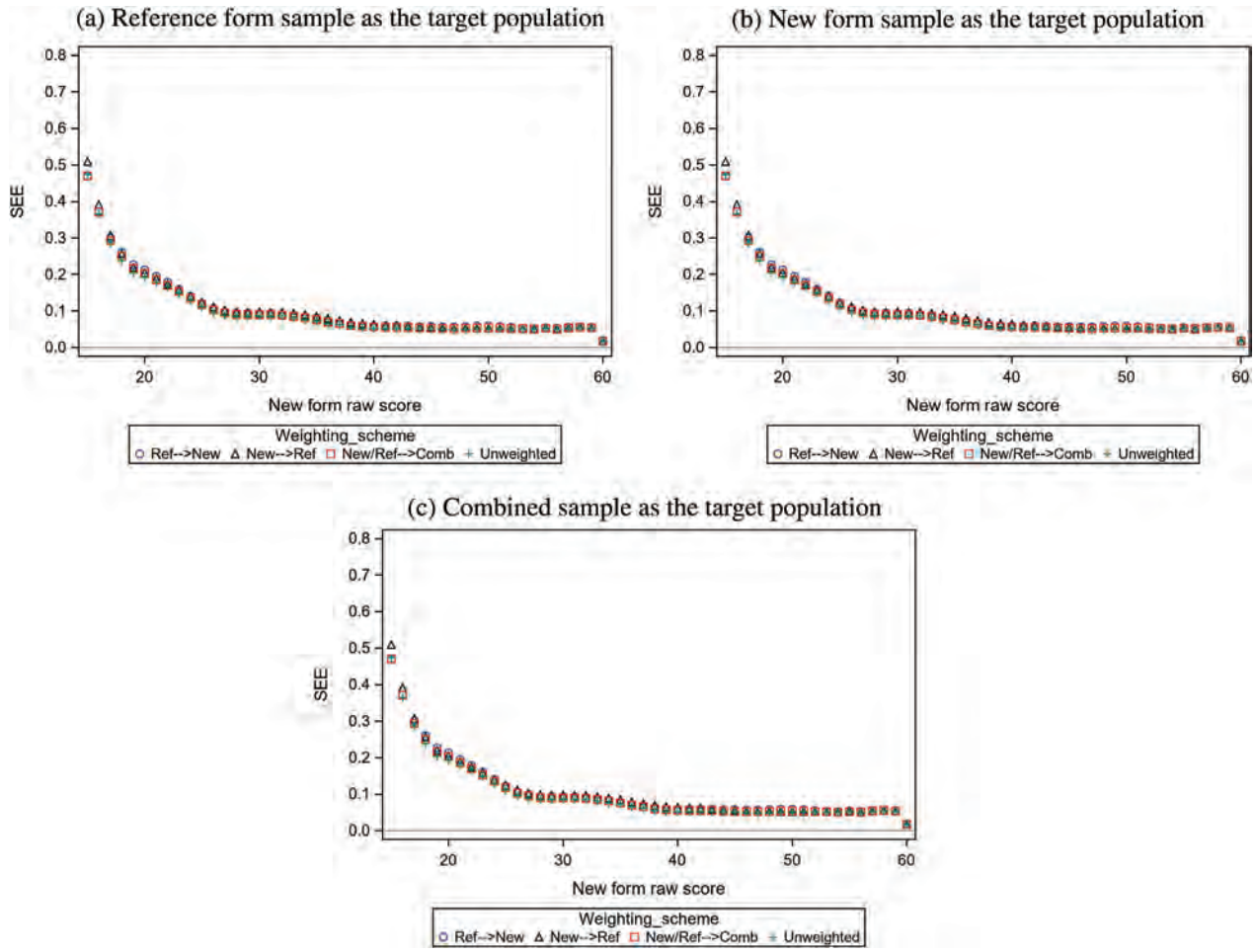

Figure 3 Conditional CSEE plots under three target population conditions.

Summary and Discussion

Score comparability of different forms of the same test is endangered in situations where the testing population is heterogenous and subgroup membership may interact with test performance. When score comparability is impacted, it affects validity issues such as population invariance and test fairness. When data are collected through the common-item design, equating can be challenging not only due to subgroup ability differences but also because the subgroups' proportions may change across the new and reference form administrations. If the equating sample is not appropriately treated, conventional equating may result in systematic bias and may have severe consequences for test takers and users.

In this study, we used operational data to create hypothetical common-item equating situations in which the groups taking the new and reference forms differed substantially in their subgroup composition. The subgroups in this study were based on real geographic regions where the testing occurred, not an artificial grouping variable created for the study. The two subgroups had a large ability difference. We specified three target populations: the new form group, the reference form group, and the combined group taking both the new and reference forms. With each target population, there was one unique criterion equating function based on the operational sample. We then carried out the conventional unweighted equating and equatings based on three different weighting schemes. One weighting scheme made the reference form sample similar to the new form sample. Another weighting scheme made the new form sample similar to the reference form sample. A third weighting scheme made both new and reference form samples similar to the combined sample.

Both the bias and RMSE measures conditioned on each raw score point and the overall average bias and RMSE measures indicated that with all three target populations, all the weighted equatings reduced the bias by at least 42%

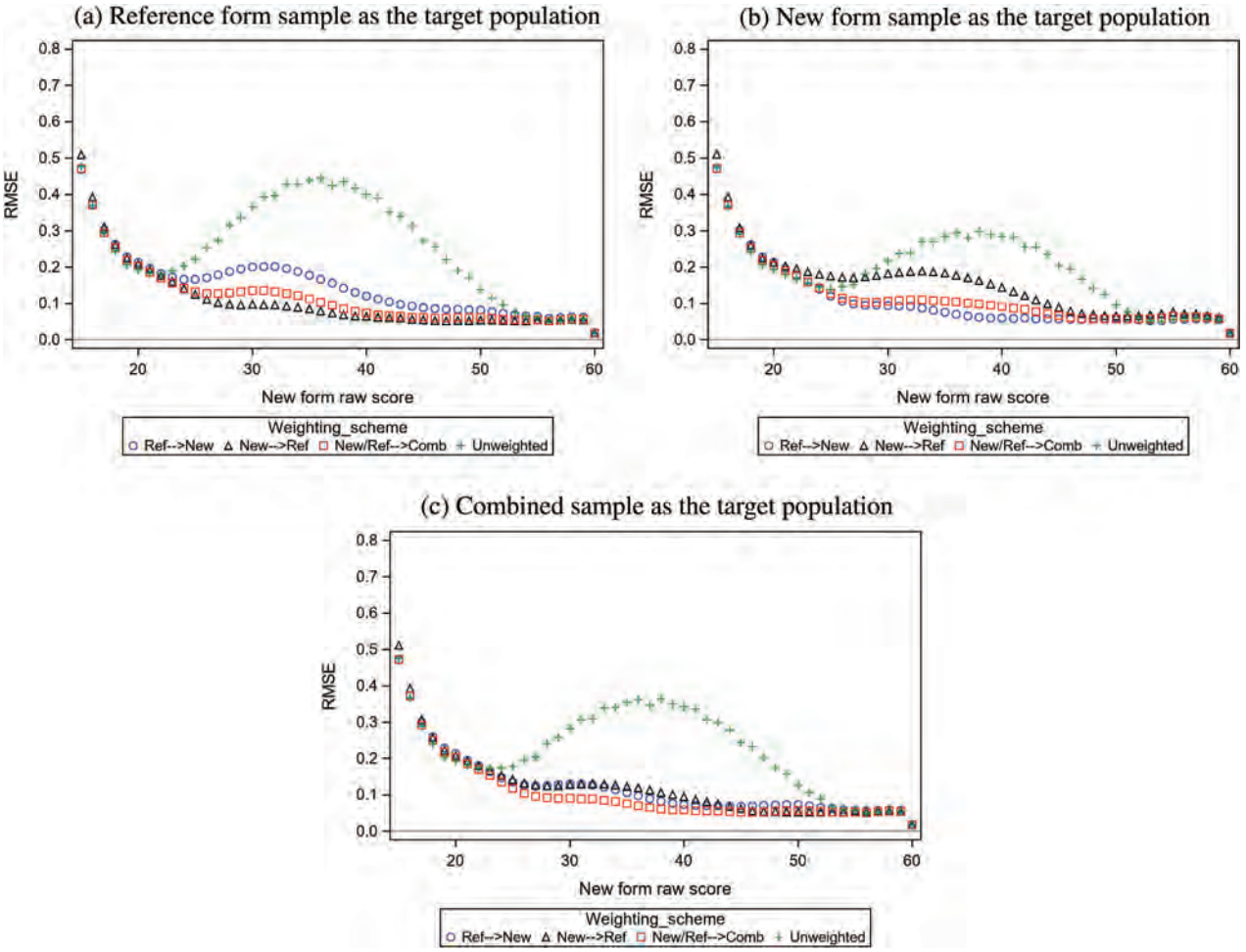

Figure 4 Conditional RMSE plots under three target population conditions.

Table 5 Average Bias Associated With Weighting Methods Used in Each Target Population Condition

	Weighting of samples			
Target population	New to reference	Reference to new	New and reference to combined	None
Reference form sample	0.01	0.11	0.05	0.29
New form sample	0.11	0.02	0.05	0.19
Combined sample	0.06	0.05	0.01	0.24

Note. New to reference = weight the new form sample to match the reference form sample; reference to new = weight the reference form sample to match the new form sample; new & reference to combined = weight both new and reference form samples to match the combined sample; none = no weighting.

((0.19 – 0.11)/0.19 = 0.42). When the weighting scheme was matched to the target population in terms of subgroup composition, the appropriate weighting reduced the bias by at least 89%. Other weighted equatings (i.e., not perfectly matched to the target population's subgroup composition) produced slightly less accurate equating results than did the weighted equating that corresponded to the subgroup composition of the target population. The results showed that equating was most accurate when the new form and reference form samples were weighted to be similar to the target population. When the target population was a combination of the two equating samples, and one sample was weighted to be similar to the other, the equating was less accurate, but still much more accurate than equating with unweighted samples.

The finding that weighted equating led to better equating results than the unweighted equating were consistent with other research on the use of common-item equating with covariates (i.e., background variables correlated with the test

Table 6 Average Standard Error of Equating (SEE) Associated With Matching Methods Used in Each Target Population Condition

	Weighting of samples			
Target population	New to reference	Reference to new	New and reference to combined	None
Reference form sample	0.10	0.09	0.09	0.09a
New form sample	0.10	0.09	0.09	0.09
Combined sample	0.10	0.09	0.09	0.09

Note. New to reference = weight the new form sample to match the reference form sample; reference to new = weight the reference form sample to match the new form sample; new & reference to combined = weight both new and reference form samples to match the combined sample; none = no weighting.

Table 7 Average Root Mean Squared Error (RMSE) Associated With Matching Methods Used in Each Target Population Condition

	Weighting of samples			
Target population	New to reference	Reference to new	New and reference to combined	None
Reference form sample	0.10	0.14	0.11	0.30
New form sample	0.14	0.10	0.11	0.21
Combined sample	0.11	0.11	0.09	0.25

Note. New to reference = weight the new form sample to match the reference form sample; reference to new = weight the reference form sample to match the new form sample; new & reference to combined = weight both new and reference form samples to match the combined sample; none = no weighting.

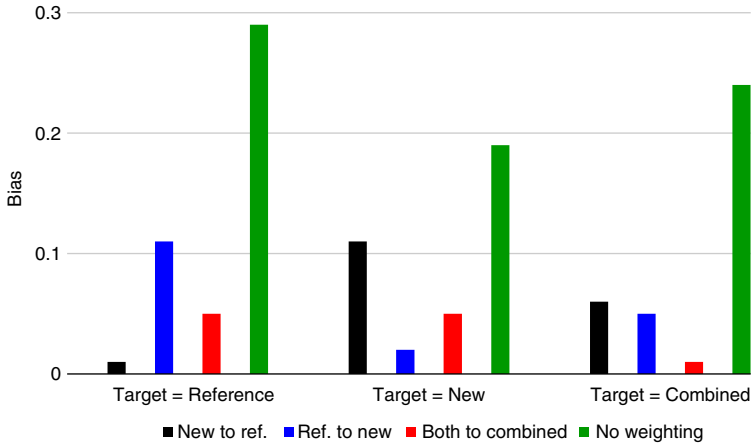

Figure 5 The average bias plot under three different target population conditions.

scores). For example, Bränberg and Wiberg (2011) showed that the use of covariates such as gender and education with linear equating had less equating error than the regular linear equating without the use of the covariates; Sansivieri and Wiberg (2017) showed reduced equating error by using covariates in the IRT observed score equating context; and Lu and Guo (2018) showed that the use of the minimum discriminant information method to create equivalent groups along with the NEAT design had less equating error than the equating without the additional background information.

Moreover, the results showed the importance for testing programs to have a specified target population in mind, even with the common item design for equating. Among the three weighting schemes, the weighted equating that matched the composition of equating samples to the target population had the least equating error across all equating conditions. In practice, even if a testing program typically has stable subgroup composition over time, the new form sample at a particular test administration could differ from the reference form sample due to a temporary event, such as the elimination or addition of testing centers in certain regions. Under this circumstance, weighting the new form sample to match the reference sample would produce the most accurate equating results. If the testing program is facing a population drift issue, weighting the reference form sample to match the new form sample would produce the most accurate equating. If

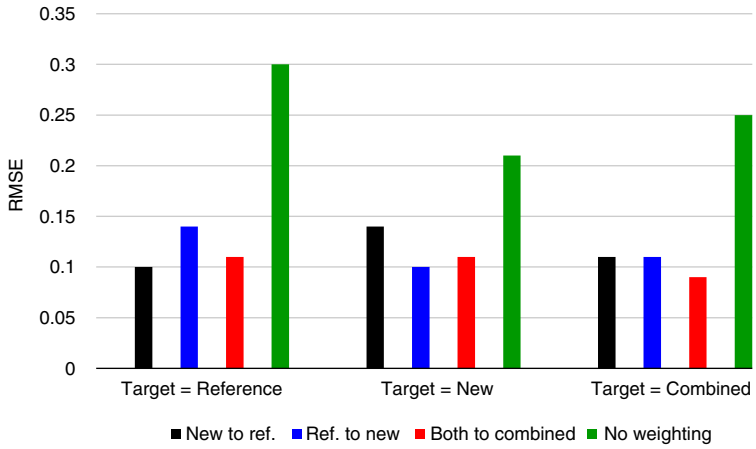

Figure 6 The average RMSE plot under three different target population conditions.

subgroup representation fluctuates across administrations, it would be desirable to use all test takers in a specified testing period to define the target population and weight both equating samples to match it at each administration.

In this study, we considered the simple case where the subgroups were characterized by only one grouping variable with two subgroups. As shown in Equation 1, weights are determined as a function of subgroup sample sizes in the new and reference forms. The proposed method could be expanded to include three or more groups classified with more than one grouping variable. With test takers' multiple background variables that are statistically related to the test scores, such as age, gender, education, and testing experiences, the weighted equating proposed in this study can still be implemented, but with a more sophisticated weighting process. Some possibilities are propensity score matching, DAG, as in Livingston (2014), the minimum discriminant information method as in Haberman (1984), or poststratification ranking as in Qian et al. (2013).

This paper examined an extreme equating case in which there were large ability differences between subgroups and substantial fluctuations in subgroup composition between the new and reference form administrations. In this case, the equating results were improved greatly by applying the proper weights to the samples. In practice, the ability differences between subgroups or subgroup composition differences across administrations may be small. Further simulation studies can be carried out to investigate the benefits of weighted equating over unweighted equating in these situations. The manipulated variables can include test length, anchor length, and sample size. Our expectation is that for programs that have multiple administrations within a year and across years, weighting equating samples can help maintain the score stability over the long run even with small fluctuations in adjacent administrations.

References

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, *48*(4), 419–440. https://doi.org/10.1111/j.1745-3984.2011.00153.x

Duong, M. Q., & von Davier, A. A. (2012). Observed-score equating with a heterogeneous target polulation. *International Journal of Testing*, *12*(3), 224–251. https://doi.org/10.1080/15305058.2011.620725

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*(3), 971–988. https://doi.org/10.1214/aos/1176346715

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772

Kim, S., & Robin, F. (2017). *An empirical investigation of the potential impact of item misfit on test scores* (Research Report No. RR-17-60). Educational Testing Service. https://doi.org/10.1002/ets2.12190

Kim, S., & Walker, M. E. (2009). *Effect of repeaters on score equating in a large-scale licensure test* (Research Report No. RR-09-27). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02184.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4310-4

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (Research Report No. RR-09-05). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02162.x

Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (Research Report No. RR-14-30). Educational Testing Service. https://doi.org/10.1002/ets2.12030

Lu, R., & Guo, H. (2018). *A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking* (Research Report No. RR-18-08). Educational Testing Service. https://doi.org/10.1002/ets2.12196

Qian, J., Jiang, Y., & Davier, A. A. (2013). *Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating* (Research Report No. RR-13-39). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02346.x

Sansivieri, V., & Wiberg, M. (2017). IRT observed-score equating with the nonequivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the psychometric society* (pp. 275–285). Springer. https://doi.org/10.1007/978-3-319-56294-0_25

Suggested citation:

Lu, R., & Kim, S. (2021). *Effect of statistically matching equating samples for common-item equating* (Research Report No. RR-21-02). Educational Testing Service. https://doi.org/10.1002/ets2.12313

Action Editor: Marna Golub-Smith

Reviewers: Gautam Puhan and Skip Livingston