Listening. Learning. Leading.®

Research Report

ETS RR–11-01

Equating of Subscores and Weighted Averages Under the NEAT Design

Sandip Sinharay

Shelby Haberman

January 2011

Equating of Subscores and Weighted Averages Under the NEAT Design

Sandip Sinharay and Shelby J. Haberman

ETS, Princeton, New Jersey

January 2011

Technical Review Editor: Matthias von Davier

Technical Reviewers: Alina von Davier and Skip Livingston

Abstract

Recently, the literature has seen increasing interest in subscores for their potential diagnostic values; for example, one study suggested the report of weighted averages of a subscore and the total score, whereas others showed, for various operational and simulated data sets, that weighted averages, as compared to subscores, lead to more accurate diagnostic information. To report weighted averages, the averages should be comparable across different test forms; that is, the averages should be equated. This report discusses how to equate weighted averages. Results from operational and simulated data sets demonstrate the small error found when equating weighted averages.

Key words: augmented subscore, chain equating, equating, weighted average

Acknowledgments

The authors thank Skip Livingston, Alina von Davier, and Matthias von Davier for useful advice, and Holly Monteith and Kim Fryer for editorial help.

The literature evidences increased interest in *subscores*, that is, scores on subtests, because they have great potential diagnostic value. The No Child Left Behind Act of 2001 requires in part that students get access to diagnostic reports developed to allow teachers to address students' specific academic needs; it is clear that subscores could be useful in the creation of diagnostic reports of this sort.

To determine the added value of subscores, Haberman (2008b) advised the use of a method with a basis in classical test theory (CTT). Haberman (2008b) also suggested that a *weighted average* of a subscore and the total score (an example of a weighted average is 0.4 × Reading subscore + 0.1 × total score) should be reported in lieu of the subscore—these averages are special cases of augmented subscores (Wainer et al., 2001). A disadvantage of weighted averages is that they can sometimes be difficult to explain; additionally, clients may balk at a reported Reading score being based not only on the observed Reading score but also on the observed Writing, Listening, and Speaking scores. However, in their research, Sinharay and Haberman (2008) and Sinharay (2010) showed that weighted averages, for a number of operational and simulated data sets, can lead to more accurate diagnostic information than subscores. For several tests, subscores do not have added value, but weighted averages do. Dwyer, Boughton, Yao, Steffen, and Lewis (2006) found that the augmented subscores (Wainer et al., 2001), which are very close to the weighted averages (Sinharay, 2010), performed noticeably well in a comparison study. Thus weighted averages are promising options for testing programs aiming to report diagnostic scores.

For weighted averages to be reported, the averages should be comparable across different forms of a given test. This comparability can be achieved by equating those forms. However, we see a dearth of research on the equating of weighted averages. For single- and equivalent-groups designs, weighted-average equating is straightforward and involves, for example, simple linear or equipercentile equating. In contrast, for *nonequivalent groups with anchor test* (NEAT) designs, graphically described in Table 1, equating weighted averages is not straightforward, and in this case, methods that perform accurately in equating the total score may not be used for equating weighted averages. One reason for this observation is that usually, only a few items in the anchor test belong to any single subscore; for

Table 1
The NEAT Design

Population	New form X	Old form Y	Anchor A
New form population P	√		√
Old form population Q		√	√

example, a test with four subscores that uses an anchor test with 24 items to equate the total score on the new form to the total score on the old form has 6 items that belong to each subscore in the anchor test. In typical cases, six items are not sufficient to allow the accurate equating of weighted averages and/or subscores. Consider also that weighted averages are often fractional in nature, but in contrast, most software packages for equating operations can only work with integer values of scores (one way around this restriction is to round the weighted averages; however, particularly for tests of minimal length, this can result in inaccurate equating). Therefore the primary objective of this report is to take a closer look at the equating of weighted averages.

There are operational tests in which scores are reported on several subareas, but the anchor test does not cover all subareas. For example, a new TOEFL® form may have some Reading and Listening items in common with an older form, but there would be no common Speaking and Writing items owing to item exposure concerns. One goal of this report, therefore, is to examine the extent of the error in equating (either systematic error or random error) the subscores and weighted averages from such a test.

In the following section, we discuss the CTT-based approach of Haberman (2008b) and the related weighted averages. The methods section begins by reviewing methods for the equating of subscores discussed by Puhan and Liang (in press); following that, the methods section covers our suggested methods for equating weighted averages. The following two sections cover the results of applying our suggested methods to two *operational* data sets, and in continuation, the next section, titled “Simulation Study,” provides the results of applying our suggested methods to several *simulated* data sets. The report then concludes, and we give some recommendations.

Weighted Averages and the Classical Test Theory–Based Approach

We denote the subscore by s and the total score of an examinee by x. Taking a CTT viewpoint, Haberman (2008b) assumed that a reported subscore is intended to be an estimate of the true subscore s_t and considered the following estimates of the true subscore:

- An estimate $s_s = \bar{s} + \alpha(s - \bar{s})$ based on the observed subscore, where $\bar{s}$ is the average subscore for the sample of examinees and α is the estimated reliability of the subscore

- An estimate $s_x = \bar{s} + c(x - \bar{x})$ based on the observed total score, where $\bar{x}$ is the average total score and c is a constant that depends on the estimated reliabilities and standard deviations of the subscore and the total score and the estimated correlations between the subscores

- An estimate $s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x})$ that is a *weighted average* of the observed subscore and the observed total score, where a and b are constants that depend on the estimated reliabilities and standard deviations of the subscore and the total score and the estimated correlations between the subscores—the *weighted average*

For a given sample, the average of the weighted averages will be the same as the average of the subscores. Therefore the weighted averages and the subscores are of the same magnitude; however, the weighted averages have smaller variance than the subscores because the weighted average corresponding to an extremely low or high subscore is pooled to the mean and is not as extreme. See the appendix for further details.

It is also possible to consider an augmented subscore s_{aug} (Wainer et al., 2001), an appropriately weighted average of all subscores of an examinee, as an estimate of the true subscore; however, the results provided by s_{aug} are very similar to the results provided by s_{sx} (Sinharay, 2010; Sinharay & Haberman, 2008). Note also that the estimate s_{sx} is a special case of the augmented subscore s_{aug}; s_{sx} places the same weight on all subscores other than the subscore of interest rather than weighing them differently.

To compare the performances of s_s, s_x, and s_{sx} as estimates of s_t, Haberman (2008b) suggested a proportional reduction in mean squared error (PRMSE), which is similar to reliability conceptually. Let us denote the PRMSEs corresponding to s_s, s_x, and s_{sx} as PRMSE$_s$, PRMSE$_x$, and PRMSE$_{sx}$, respectively; the larger the PRMSE, the more accurate is the estimate. For a subscore to have added value, PRMSE$_s$ has to be larger than PRMSE$_x$ (Haberman, 2008b). Haberman (2008b) also recommended that for the weighted average to have added value, PRMSE$_{sx}$ has to be substantially larger than both PRMSE$_s$ and PRMSE$_x$. (The appendix provides more details on PRMSEs.)

Sinharay and Haberman (2008) and Sinharay (2010) showed that for several operational and simulated data sets, weighted averages, compared to subscores, lead to more accurate diagnostic information. For example, these authors demonstrated that subscores do not have added value, but weighted averages do, for several tests; additionally, Sinharay (2010) demonstrated that weighted averages typically perform nearly as well as augmented subscores (Wainer et al., 2001), and so for testing programs interested in reporting diagnostic scores, weighted averages are promising.

However, though Puhan and Liang (in press) suggested several methods for equating subscores, no techniques for equating weighted averages exist for the NEAT design. Equating of weighted averages is covered in the following section.

Methods

In this section, we discuss methods suggested by Puhan and Liang (in press) for the equating of subscores, facilitating discussion of methods for the equating of weighted averages, covered next.

Methods for Equating Subscores

Puhan and Liang (in press) suggested two approaches for equating subscores, both making use of linear and nonlinear equating methods. Here we focus only on the nonlinear equating methods. We denote as A the anchor test (external or internal) used to equate the total score of a test. The first method equates a subscore in a new form (X) to the corresponding subscore in an old form (Y) using as an anchor score the score on the items that are in A and that belong to the same subscore (Puhan & Liang, in press). The method

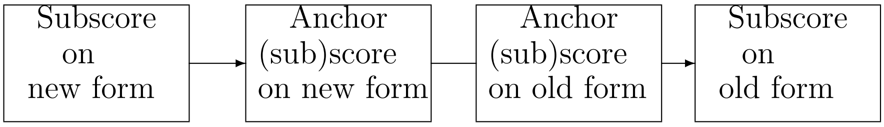

Figure 1. **The first method of equating subscores shown graphically.**

is illustrated graphically in Figure 1.

In Figure 1, the score on the items that are in A and belong to a subscore are referred to as the *anchor (sub)score*. A one-sided arrow indicates a single group equating of the score adjacent to the source of the arrow to the score adjacent to the destination of the arrow. A line indicates that the scores connected by the line are on the same scale.

The second method proceeds in a similar manner. The total score (i.e., the sum of all subscores) on the new form is equated first to the total score on the old form. Then this method uses the total score on the old form as an anchor score for the old form population and, for the new form population, the equated total score on the new form. Thus it can be seen, using this method, that one transforms, using a single-group equipercentile equating on the new form population, a subscore S_N on the new form to a equated total score $T(S_N)$ on the new form. Because the latter is on the scale of the total score on the old form, one then transforms $T(S_N)$ to a subscore $S_O(T(S_N))$ on the old form using a single-group equipercentile equating on the old form population. The method is illustrated graphically in Figure 2.

It has been found (Puhan & Liang, in press) that when the proportion of a subtest common between the old and new form is small, the second method performs better than the first, and vice versa. The same authors found that equating the subscores using one of these methods is better than not equating the subscores in terms of producing scores that are more fair.

Puhan and Liang (in press) also referred to a third method whereby the total score

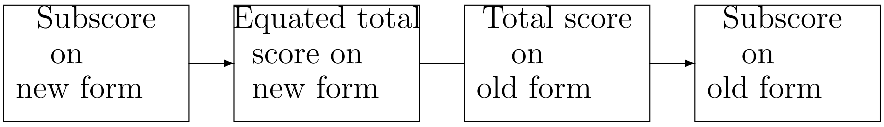

Figure 2. **The second method of equating subscores shown graphically.**

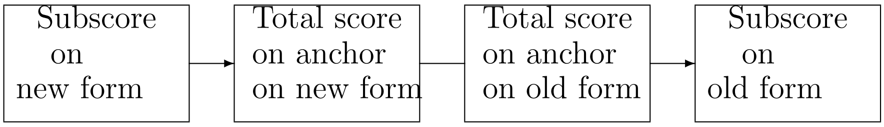
Subscore on new form
Total score on anchor on new form
Total score on anchor on old form
Subscore on old form

Figure 3. **The third method of equating subscores shown graphically.**

on the anchor test is used as an anchor score to equate the subscores. The method is illustrated graphically in Figure 3.

The third method performs in a similar way to the second and so is not discussed further in this report. Note as well that the second and third methods can be employed to equate a subscore that is not represented in the anchor test.[1]

Methods for Equating Weighted Averages

Here we discuss our suggested methods for equating weighted averages under the NEAT design. All the methods use some kind of anchor score to do what is essentially the chain equipercentile equating (Kolen & Brennan, 2004). Here the anchor score is actually a score that is on the same scale for both the new form and old form population; it allows one to adjust for the difference in difficulty of the two test forms. The weighted average on the current form is equated to the anchor score for the new form population, which is on the same scale as the anchor score for the old form population, and then the anchor

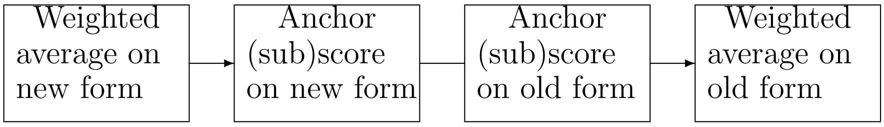
Weighted average on new form
Anchor (sub)score on new form
Anchor (sub)score on old form
Weighted average on old form

Figure 4. **The first method of equating weighted averages shown graphically.**

score on the old form population is equated to the weighted average on the old form. The methods only differ in the way in which the anchor scores are defined.

The first method proceeds like the first method of Puhan and Liang (in press) by equating a weighted average on a new form to the corresponding weighted average on an old form and using the anchor (sub)score as an anchor score. This is illustrated graphically in Figure 4.

The reader should note that in this method as well as in the following methods, the weights on the weighted averages that are being equated are determined by the formula from Haberman (2008b); consequently, the weights between the old and new forms differ (i.e., we may be equating 0.4 × Reading subscore + 0.1 × total score on the new form to 0.3 × Reading subscore + 0.16 × total score on the old form). It is therefore possible to (a) compute weighted averages on several forms of a test and then, if the weights are very close (or the weighted averages based on different weights are highly correlated), to (b) fix the weights at a specific set of values (e.g., at the average of the weights) and (c) equate the weighted averages with fixed weights between forms (i.e., equate 0.35 × Reading subscore + 0.13 × total score on the new form to 0.35 × Reading subscore + 0.13 × total score on the old form). However, in the remainder of this report, we do not consider equating of weighted averages with fixed weights. Weighted averages with varying weights are the regression estimates of the corresponding true subscores, and equating them makes more sense (i.e., we are essentially equating the regression estimate of true subscores in the new form to the regression estimate of true subscores in the old form). Weighted averages with

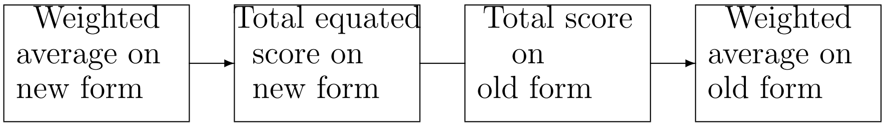
Weighted average on new form
Total equated score on new form
Total score on old form
Weighted average on old form

Figure 5. **The second method of equating weighted averages shown graphically.**

fixed weights, though convenient, will not have this property.

The second method of equating weighted averages is similar to the second method of Puhan and Liang (in press) for equating subscores: The total score (i.e., the sum of all subscores) on the new form is first equated to the total score on the old form. Following that, the investigator uses as an anchor score the total score on the old form for the old form population and the total equated score on the new form for the new form population. This method is illustrated graphically in Figure 5.

The third method uses as an anchor score a weighted average of the scores on the subareas on the anchor test. This is illustrated graphically in Figure 6. Any set of weights can be employed for computation of the anchor scores. One can fix the weights based on experience obtained from previous administrations of the test or estimate the weights from the current sample. Here we employ the average of the weights on the subscores from the new form and old form populations. As an example, consider that the interest here is in equating the weighted average corresponding to Subtest 1. In that case, one first computes the weights on Subscore 1 and the total score in Weighted Average 1 in the new and old form populations and then averages these values. The average weights are then imposed on the anchor (sub)scores to compute an anchor score to equate Weighted Average 1. We plan to explore other weights in future research.

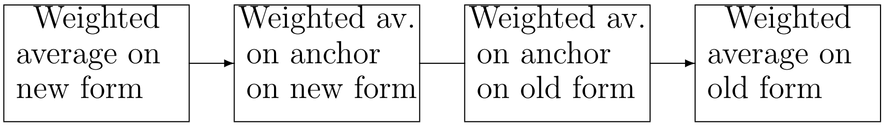
Weighted average on new form
Weighted av. on anchor on new form
Weighted av. on anchor on old form
Weighted average on old form

Figure 6. **The third method of equating weighted averages shown graphically.**

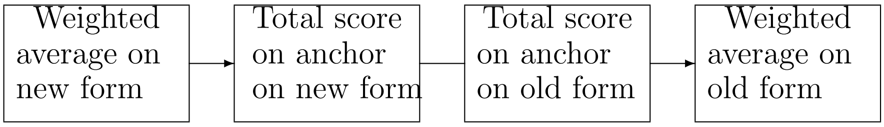
Weighted average on new form
Total score on anchor on new form
Total score on anchor on old form
Weighted average on old form

Figure 7. **The fourth method of equating weighted averages shown graphically.**

The fourth method is similar to the second method and uses the total score on the anchor test as an anchor score to equate the weighted averages. This method is illustrated graphically in Figure 7. Because this method is performed exactly like the second method, this report does not discuss any further results on this method. Also, it should be noted that the previous three methods can be employed to equate a subscore that is not represented in the anchor test.[2]

The last two methods of equating subscores and the last three methods of equating weighted averages use, as anchor scores, not only the corresponding subscore but also the other subscores. For example, in the equating of a Reading subscore or weighted average for reading, scores on mathematics, science, or writing items may also contribute to the anchor score. Though this may seem counterintuitive, it leads to sufficiently accurate results because subscores are most often highly correlated with each other in operational tests (see, e.g., Sinharay, 2010). Additionally, these methods show some similarities to the equating of AP$^{®}$ examinations, incorporating both multiple choice (MC) and constructed response (CR) items. In these examinations, the MC score on a new form is equated to the MC score on an old form using an MC anchor test, and because the MC score on the old form has been equated to the MC score on a base form, it is equated to the MC score on a base form. Following this, the composite score on the new form, a weighted average of the MC and CR scores, is linked to the MC score on the new form using single-group linking. Together these equatings create a link from the MC score on the base form to the composite score on the new form. The final link, used later to convert cut scores from the scale of MC scores on the base form to the scale of composite scores on the new form, is accurate owing to the high correlation between the MC and composite scores (see, e.g., Sinharay & Holland, 2007). In this report, the correlation between the score to be equated and the anchor score is expected to be moderate for the last two methods of equating subscores and high for the last three methods of equating weighted averages. For this reason, we expect these methods to lead to accurate equating.

Additionally, the literature has suggested using scores on MC items as anchors for equating for several tests involving CR items (Ercikan et al., 1998; ETS, 2007), even though it is believed that CR and MC items often measure slightly different skills. In a study using data from several operational tests, Livingston (1994) found that for tests with only a few CR items, it is preferable to use a related test with MC items as an anchor for equating.

In the following section, we show the results of the performance of the previously mentioned methods in two applications to two operational data sets. We have as an objective the examination of the accuracy of the methods and a comparison of the methods.

Application 1

Data

The original data for this example are from one form of a licensing test for prospective teachers. The test form included 119 MC items, approximately equally divided among four content areas, including mathematics, language arts, science, and social studies. Scores on each of these content areas are reported; these are treated as the subscores in this report. Previously, the original form had been used at two test administrations; the two examinee populations are represented by P and Q in this analysis.

The item responses from the original test were used to construct two pseudotests: X (new form) and Y (old form). A pseudotest consists of a subset of the test items from the original 119-item test, and the score on the pseudotest for an examinee is found from the responses of that examinee to the items in the pseudotest. Each of Pseudotests X and Y contains 44 items, including 11 items from each of the four content areas mentioned earlier. Pseudotests X and Y had no items in common and were adapted to be parallel in content. A set of 24 items (6 from each content area) was selected to be representative of the original test and to serve as the external anchor A, which had no items in common with either Pseudotest X or Pseudotest Y. The mean percentage correct on the anchor test approximately equaled that for the 119-item original test. One can find further details on the construction of these pseudotests in the work of von Davier et al. (2006) and Holland, Sinharay, von Davier, and Han (2008).

Pseudotest X was constructed to be considerably easier than Pseudotest Y; for example, on Q, the mean score for X is larger than the mean score for Y by 133% of the standard deviation of Y. In addition, Q is more able than P with a mean A score that is higher than P by approximately a quarter of a standard deviation in $P + Q$.

These pseudotests were designed to produce an equating problem for which solutions would be nonlinear and the different equating methods would be expected to give different results. The great difference in difficulty between Pseudotest X and Pseudotest Y ensured that the equating functions would be nonlinear and the relatively large difference in test performance between P and Q ensured that different equating methods would produce different results.

For this test, four subscores, one corresponding to one content area, are reported operationally. Figure 8 shows the average proportion of correct scores of P and Q on the four subtests and the total test for the original 119-item test. The figure shows that the difference between the two populations is mostly similar across the four subscores, and this finding is true even if the subscores belong to a variety of content areas.

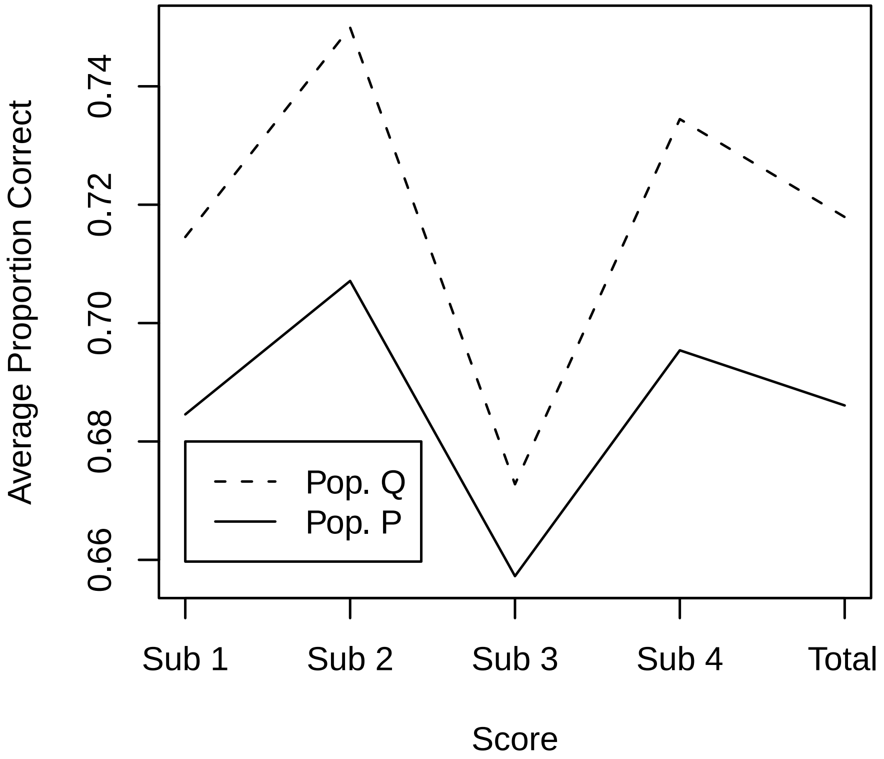

Figure 8. **The differences in the proportion of correct scores between P and Q for Application 1.**

Table 2
Proportional Reduction in Mean Squared Errors of the Subscores for Application 1

Subscore	PRMSE$_s$	PRMSE$_x$	PRMSE$_{sx}$
1	0.57	0.75	0.77
2	0.62	0.68	0.73
3	0.49	0.67	0.69
4	0.62	0.75	0.78

Note. PRMSE = proportional reduction in mean squared error.

Table 2 provides the values of the PRMSEs computed from the data on X in P. The PRMSEs are very similar if they are computed on Y instead of X or on Q instead of P, or both. The reliability of the scores on X in P is .82. Table 2 shows that though the subscores do not have added value, the weighted averages do, and thus this data set is seen as appropriate for applying the previously mentioned methods for the equating of weighted averages.

Analyses and Results

As all examinees for P and Q took all 119 items on the original test, all the examinees for P and Q have scores on X, Y, and A. To mimic the structure of the NEAT design, in this study, we ignore the scores on X in Q and on Y in P and perform the equatings described previously. We also consider that because the data usually missing in the NEAT design are in fact available for the pseudotest data, there is a possibility to compute the true equating of the subscores and of weighted averages by performing a single-group equating of the subscores or weighted averages on X to the corresponding quantities on Y using the group $P + Q$.

We used the two subscore equating methods described in the methods section in equating the subscores using these data. These data were also used to equate the weighted averages using the three methods for equating weighted averages discussed in the methods section. Consider also that the weighted averages can have fractional values, for example, the possible number of weighted averages corresponding to Subscore 1 on X in P was found to be 408 for the data set: They are 2.43, 2.59, …, 10.79, 10.95. Therefore it is possible to compute a function equating any real value between 0 (minimum possible value on the subtest) and 11 (maximum possible value on the subtest) of a weighted average on a subtest on X to a weighted average on a subtest on Y. By way of simplification, here we compute the function equating weighted averages only at integer values between 0 and 11 (and we note that the weighted averages on X in P or Y in Q were not rounded before the equating). We treated the anchor test as external, although the methods discussed in this report apply also to internal anchor tests. We used polynomial loglinear models (Holland & Thayer, 2000) to presmooth the joint and conditional distributions of subscores and total score, the linear interpolation method (Kolen & Brennan, 2004) to continuize the discrete subscore distributions during the equating, and a version of the linear interpolation method appropriate for fractional scores to continuize the discrete distributions of weighted averages during the equating.

We show in Figure 9, for each content area, the equating functions for the subscores and weighted averages for Method 2, which uses the total test score as an anchor score. The 45° line shows identity equating or no equating,[3] and we see a substantial difference between the identity equating and equating functions for Subscores 1, 2, and 4.

We show with Figure 10, for each method, the differences in the equating and true equating functions for each subscore. The figure also shows the 5th and 95th percentiles of the subscores in the combined population using vertical lines. In each panel of the figure, the range of the X axis is the 1st and 99th percentile of the corresponding subscore in P. Furthermore, as a baseline for comparison, differences for no equating are also added. The standard deviation of the subscores on X range between 2.5 and 3.1 for the four subtests.

For each subscore and method, we computed an overall measure of difference between an observed equating function and the true equating function by computing the square root of the weighted average of the square difference between the observed and true equating functions, where the weight at a subscore point is proportional to the frequency of that subscore on X in $P + Q$. If the observed equating function, the true equating function, and the weight at subscore point i are denoted by o_i, t_i, and w_i (where $\sum_i w_i = 0$),

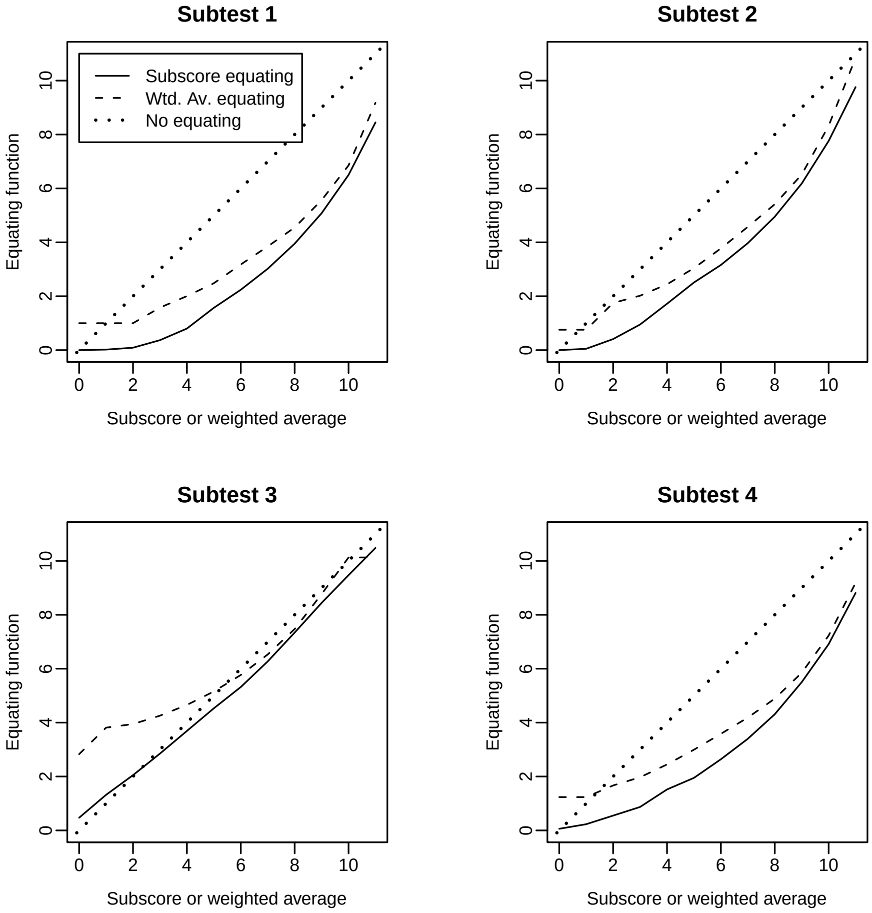

Figure 9. **The equating functions for the subscores and weighted averages for Application 1.**

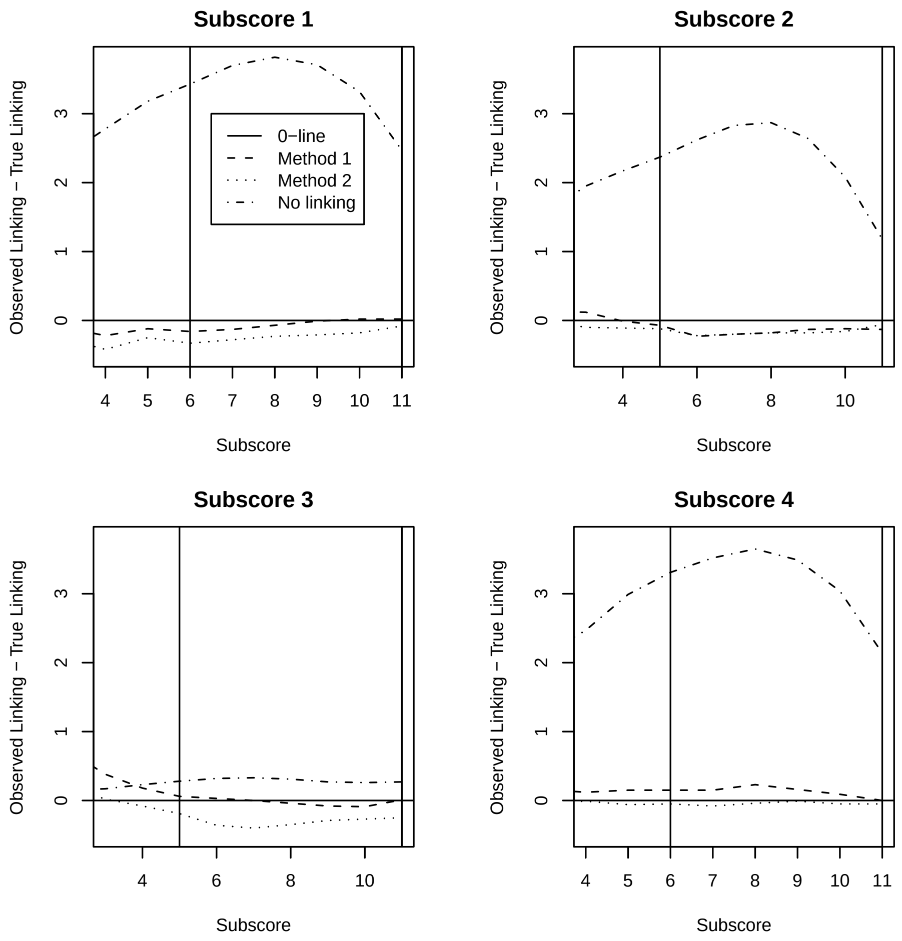

Figure 10. **The differences between the observed and true subscore equating functions for Application 1.**

respectively, then the measure is given by $\sqrt{\sum_i w_i(o_i - t_i)^2}$. We refer to this measure as the root mean squared difference (RMSD). The values of the measure for the four subscores and the two methods and no equating are given in Columns 2–4 of Table 3.

Table 3
Comparative Performance of the Methods for Equating Subscores and Weighted Averages for Application 1

	Subscore equating			Equating of weighted average			
Content area	Method 1	Method 2	No equating	Method 1	Method 2	Method 3	No equating
1	.06	.19	3.25	0.15	0.08	0.09	3.22
2	.14	.15	2.23	0.14	0.07	0.06	2.17
3	.10	.32	0.29	0.14	0.09	0.08	0.38
4	.12	.05	3.00	0.26	0.16	0.16	2.93

Figure 10 and Columns 2–4 of Table 3 show that there is a substantial difference between Method 1 and Method 2 for equating subscores: Specifically, Method 1 performs better (in terms of having lower RMSD) for the first three subscores, whereas Method 2 performs better for the fourth subscore. Versus no equating, any of the methods for equating is better. In addition, the RMSD in any of these methods is small; for example, the RMSD values previously computed are in units of raw subscore points, where a difference of 0.5 or more is usually a difference that matters (DTM). In other words, only a difference more than a DTM usually leads to different equated raw scores (Dorans & Feigenbaum, 1994). The RMSD values of Table 3 are all less than 0.5; that is, they are all less than the DTM.

For each of the three methods, Figure 11 shows the differences of the equatings of the weighted averages and the true equating function. The figure also shows, using vertical lines, the 5th and 95th percentiles of the weighted averages in the combined population. The standard deviations of the weighted averages on X range between 1.8 and 2.6 for the four subtests.

The last four columns of Table 3 show the values of a measure similar to that reported in Columns 2–4 for the four weighted averages. To compute the weights w_is, we

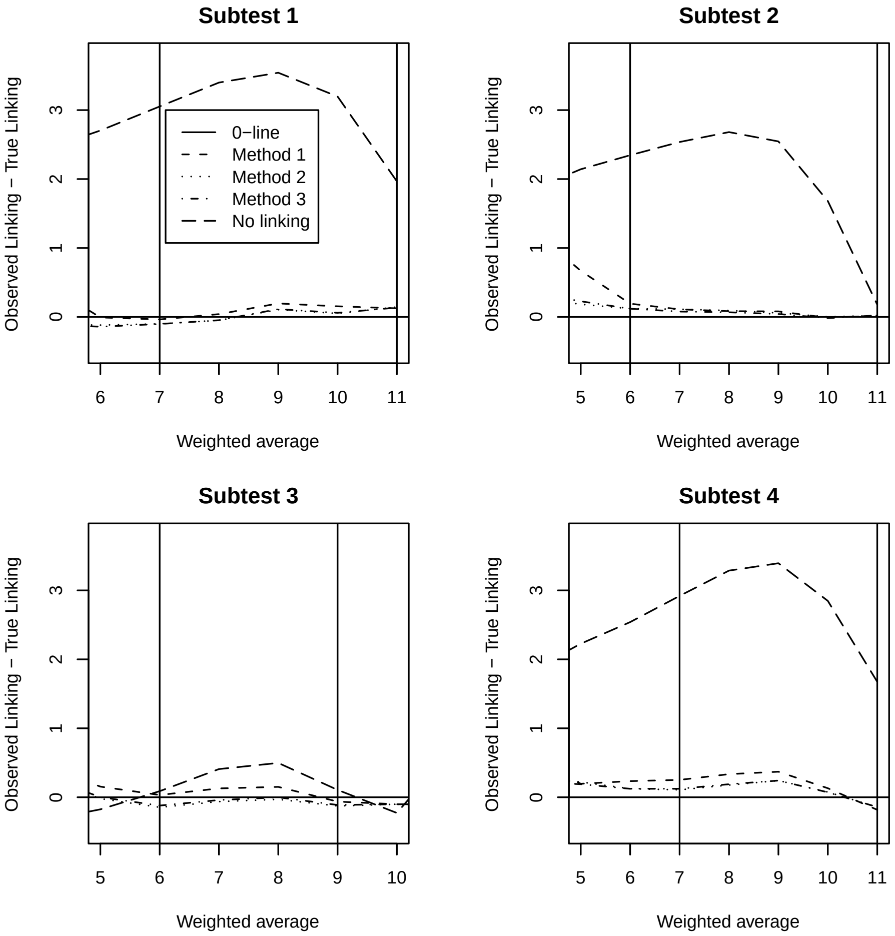

Figure 11. **The differences between the observed and true equating functions for equating weighted averages for Application 1.**

rounded the values of the weighted averages on X in $P + Q$ to the nearest integers and computed the weights as proportional to the frequency of a rounded weighted average.[4]

Figure 11 and the last three columns of Table 3 show that Methods 2 and 3 for equating weighted averages perform very similarly and better than Method 1. As with the subscores, any method for equating weighted averages is better than no equating at all.

Columns 2–4 of Table 4 show the correlation between the subscore and the anchor score for population Q for the two methods for equating subscores. The last three columns of Table 4 show the correlation between the weighted average and the anchor score for population Q for the three methods of equating weighted averages.

Table 4
Correlation of the Subscore and the Anchor Score and Weighted Average and Anchor Score in Q

Content	Correlations for subscores		Correlations for weighted averages		
area	Method 1	Method 2	Method 1	Method 2	Method 3
1	0.46	0.75	0.55	0.99	0.73
2	0.64	0.78	0.69	0.96	0.76
3	0.37	0.70	0.45	0.98	0.71
4	0.46	0.78	0.52	0.99	0.74

In Table 4, we associate Method 2 with highest correlations for both subscores and weighted averages. Thus we are not surprised by the good performance of Method 2 for equating weighted averages. However, the good performance of Method 1 for equating subscores compared to Method 2 is slightly unexpected given these correlations, which indicates that we require more than high anchor score–total score correlation for accurate equating.

Analyses and Results for the Case When the Anchor Test Is Nonrepresentative

In this analysis, we use the same Pseudotests X and Y as in the earlier analysis; however, we use a shorter 12-item anchor test $A1$ that has six items each from the first two content areas (the same items as in A) and no items belonging to the last two content areas, instead of the 24-item anchor test A mentioned earlier; that is, the anchor test is nonrepresentative of the tests to be equated in this situation. We use Method 2 to equate the subscores and Method 2 for equating weighted averages. Note that Method 1 for equating either the subscores or the weighted averages cannot be used for the last two content areas, and so we did not do so in this study. Table 5 shows the RMSD values for the equatings of subscores and weighted averages.

Note that even in this case, the RMSD for no equating is given by the corresponding numbers in Table 3, and also, equating is seen to be much better than no equating at all. Table 5 shows that the equating methods lead to accurate results even in the case considered here. The RMSD values are only slightly worse than the values of Table 3. In addition, the RMSD values for the last two content areas are small, despite that the anchor test does not contain any items belonging to these two content areas. This is possible because the total scaled scores perform well as anchor scores; in other words, the difference in the average total scaled score between P and Q reflects the difference in any average subscore between P and Q (as evident from Figure 8). The length of the anchor test (of 12 items) in this case is less than the length recommended by most equating experts (see, e.g., Kolen & Brennan, 2004), but, similar to earlier, equating with such a short anchor test is better than no equating at all.

Table 5
Comparative Performance of the Three Methods for Equating Weighted Averages in Application 1 When the Anchor Is Nonrepresentative

Content area	Subscore equating: Method 2	Equating weighted averages: Method 2
1	0.19	0.08
2	0.16	0.07
3	0.34	0.11
4	0.07	0.18

Application 2

Data

The pseudotests in this example were constructed in a similar way to how they were constructed in the first example and use data from two administrations of a different form of the same testing program (data used by Puhan, Moses, Grant, & McHale, 2009). This form had 120 MC items. Two Pseudotests X and Y, both with 48 items, and an external anchor A, with 24 items, were created from the original test. As earlier, this anchor had no items in common with either Pseudotest X or Pseudotest Y.

This example is less extreme than the example given earlier. Though the pseudotests are substantially different in terms of difficulty, with Pseudotest X being harder than Pseudotest Y, the difference is not as large as it was in the preceding section; for example, in Q, the mean score on Pseudotest X is smaller than the mean score on Pseudotest Y, but only by about 43% of the standard deviation of Y. In the same way, Q is more able than P and has a mean A score that is higher than the A score in P by only 14% of the standard deviation in $P + Q$. Nevertheless, these differences were expected to be large enough to lend the equating functions a significant nonlinear component and to cause various equating methods to differ. As in the preceding operational data example, only one of the four subscores has added value, and all the weighted averages have added value here.

Analyses and Results

Figure 12 shows, for each of the three methods, the differences of the equatings of the weighted averages and the true equating function. Table 6 shows the root mean squared error (RMSE) values described previously. As in our first application, Figure 12 and Table 6 show that Method 2 performs differently from Method 1 for both subscores and weighted averages. Similar to the preceding application, Methods 2 and 3 for equating weighted averages perform mostly better than Method 1, and equating using any method is better than no equating.

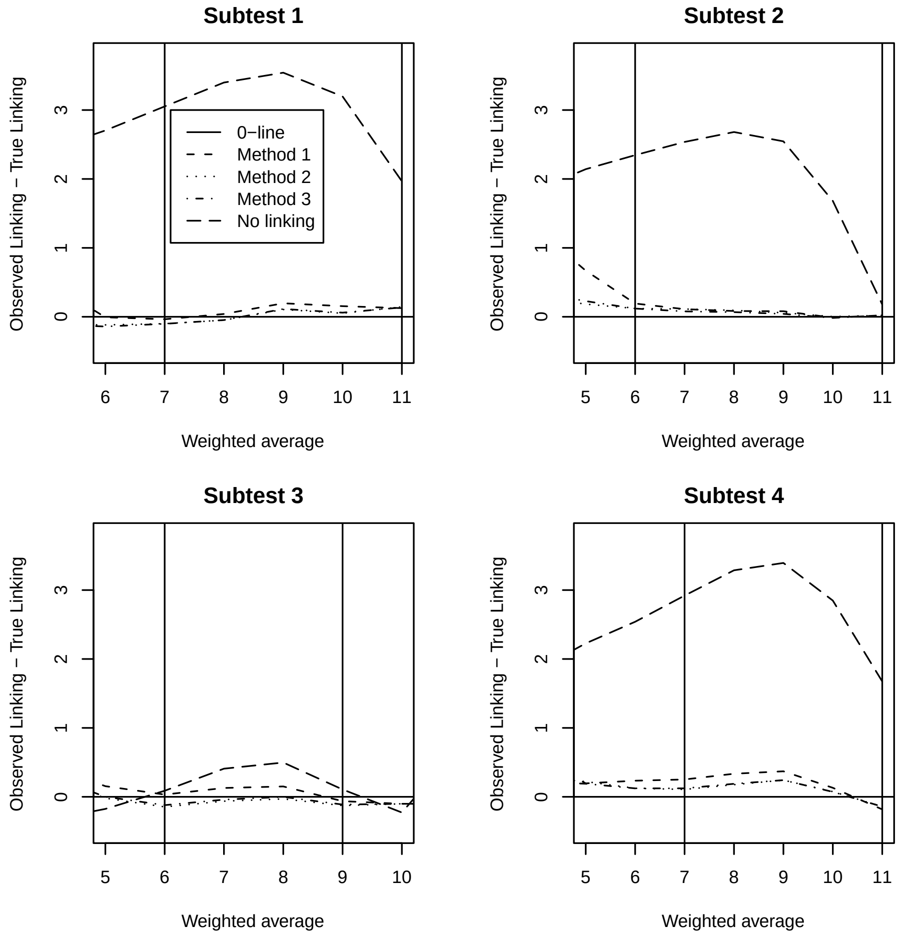

Figure 12. **The differences between the observed and true equating functions for equating weighted averages for Application 2.**

Table 6
Comparative Performance of the Methods for Equating Subscores and Weighted Averages for Application 2

	Subscore equating			Equating of weighted average			
Content area	Method 1	Method 2	No equating	Method 1	Method 2	Method 3	No equating
1	0.08	0.09	1.13	0.15	0.11	0.11	1.18
2	0.10	0.17	0.54	0.21	0.26	0.19	0.56
3	0.04	0.16	1.12	0.17	0.11	0.11	1.08
4	0.19	0.15	1.29	0.19	0.05	0.06	1.34

Table 7 shows the correlation between the weighted average and the anchor score for population Q for the three methods of equating weighted averages. The correlations in Table 7 are mostly higher than those in Table 4.

Table 7
Correlation of the Weighted Average and the Anchor Score for Q

Content area	Method 1	Method 2	Method 3
1	0.81	0.99	0.96
2	0.73	0.99	0.96
3	0.77	0.99	0.95
4	0.78	0.99	0.96

Analyses and Results for the Case When the Anchor Test Is Nonrepresentative

Similar to Application 1, we used a shorter 12-item anchor test $A1$ comprising six items each from the first two content areas (the same items as in A) and with no items from the last two content areas. We used Method 2 for equating subscores and Method 2 for equating weighted averages. Table 8 shows the RMSE values for the equatings of subscores and weighted averages for this case.

Similar to the preceding application, the RMSE of equating the subscores and weighted averages is small even for the content areas that do not contribute any items to the anchor test. Also, even though the anchor is shorter than what is recommended by experts, equating using it is better than no equating.

Simulation Study

Though the two operational data examples have given us some insight into the performance of our suggested methods, the examples represent only a small fraction of all possible equating scenarios. For example, they do not involve long subtests. For this reason, we performed a simulation study to examine the performance of the earlier equating methods under several more scenarios; the study is similar to the studies reported by Sinharay and Holland (2007) and Sinharay (2010).

Simulation Design

We obtained a data set from a licensing test for prospective teachers. The 118 MC items on the test belonged to four subscores: Items 1–29 were on language arts, Items 30–59 were on mathematics, Items 60–88 were on social studies, and Items 89–118 were on science. Because the subscores measure four different but correlated dimensions, we fitted a multidimensional item response theory (IRT) model (MIRT; e.g., Reckase, 2007) with a response function (for item i):

$$\frac{e^{a_{1i}\theta_1+a_{2i}\theta_2+a_{3i}\theta_3+a_{4i}\theta_4-b_i}}{\left(1+e^{a_{1i}\theta_1+a_{2i}\theta_2+a_{3i}\theta_3+a_{4i}\theta_4-b_i}\right)}, \boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3,\theta_4)' \sim \mathcal{N}_4\left(\boldsymbol{\mu}=(0,0,0,0)',\Sigma\right), \tag{1}$$

where a_{ji} are slope parameters and b_i are location parameters. Each component of $\boldsymbol{\theta}$ belongs to an operational subscore. The diagonals of Σ are set to 1 to ensure identifiability of the model parameters. For item i, only one of either a_{1i}, a_{2i}, a_{3i}, or a_{4i} is assumed to be nonzero, depending on the item content (e.g., for an item from the first content area, a_{1i} is nonzero, whereas $a_{2i}=a_{3i}=a_{4i}=0$) so that the MIRT model has a simple structure.

The estimated item parameter values from the fitting of the model given in Equation (1) to the data set were instrumental in obtaining the generating item parameters of Pseudotests X and Y in the simulations. A bivariate normal distribution $\mathcal{D}_k$ was fitted to the log-slope and difficulty parameter estimates corresponding to the kth content area, $k=1,2,3,4$. The generating item parameters for the kth content area of Y were randomly drawn from $\mathcal{D}_k$. To compute the generating item parameters for the kth content area of X, we made random draws from $\mathcal{D}_k$ and then added a constant $\Delta_d=0.25$ to the

Table 8
Performance of the Methods for Equating Subscores and Weighted Averages in Application 2 When the Anchor Is Nonrepresentative

Content area	Subscore equating: Method 2	Equating of weighted average: Method 2
1	0.12	0.10
2	0.08	0.16
3	0.22	0.09
4	0.19	0.07

difficulty parameter component of all the draws. This strategy ensured that Pseudotest X was slightly more difficult than Pseudotest Y, and the difference was similar over all subscores. Because we are generating data from a MIRT model, Pseudotest X can differ from Pseudotest Y in other ways (see, e.g., Sinharay & Holland, 2007), but this report does not consider those ways.

We assume that the length of the anchor test is about 40% of that of Pseudotest X or Y. To compute the generating item parameters for the kth content area for the anchor test A, we made random draws from the distribution $\mathcal{D}_k$ and then added a constant $\Delta_d/2$ to the difficulty parameter component of all the draws. This strategy ensured that the difficulty of A was between the Pseudotest X and Y difficulties. The generating item parameters for Pseudotest X, Y, and A were identical for all R replications in a given simulation condition.

Factors controlled in the simulation. The following two factors were controlled in the simulation:

1. *Length of the subscores.* This report used three length values: 12, 20, and 30. The reliability of a test increases as the test length increases. For simplicity, this report assumes that the different subscores for a given test have the same length.

2. *Level of correlation (ρ) among the components of $\boldsymbol{\theta}$.* This report used three levels of .70, .80, and .90; a survey of operational data (Sinharay, 2010) showed this to be a realistic range. If the correlation level for a simulation case is ρ, the mean of all the off-diagonal elements of Σ (denoting the correlations between the components of $\boldsymbol{\theta}$)

in Equation (1) was set equal to ρ to simulate the data sets. The starting point in obtaining such a Σ was

$$\mathcal{C} = \begin{pmatrix} 1.00 & .78 & .80 & .84 \\ & 1.00 & .72 & .78 \\ & & 1.00 & .88 \\ & & & 1.00 \end{pmatrix}.$$

This is the estimated correlation matrix $\mathcal{C}$ between the components of $\boldsymbol{\theta}$ from the fit of the model given by Equation (1) to the previously mentioned data set from the licensing test. To obtain a Σ with a mean correlation ρ, we computed the mean of the correlations of $\mathcal{C}$, m. Afterward, the (i, j)th element of Σ was set as the (i, j)th element of $\mathcal{C} - m + \rho$, where $i \neq j$.[5] Through this strategy, we were able to ensure that the average of the correlations in Σ was ρ, but the strategy also allowed the correlations between the subscores to be different in a realistic way.

The steps of the simulation. For each simulation condition (determined by a *correlation* and a *subscore length*), the generating item parameters of Pseudotest X, Pseudotest Y, and the anchor test A were randomly drawn (as we described previously) once. Afterward, we performed the $R = 100$ replications. The sample size of both P and Q was set to 2000. Each replication involved the following steps: (a) We generated the ability parameters, (b) simulated the scores, and (c) performed equating. Step details are provided in the following:

1. Generate the ability parameters $\boldsymbol{\theta}$ for the populations P and Q from ability distributions $g_P(\boldsymbol{\theta})$ and $g_Q(\boldsymbol{\theta})$, respectively. We used $g_Q(\boldsymbol{\theta}) = \mathcal{N}_4(\mathbf{0}, \Sigma)$, where Σ is obtained as described earlier, to ensure that the average correlation is ρ. We used $g_P(\boldsymbol{\theta}) = \mathcal{N}_4(\boldsymbol{\mu}_P, \widehat{\Sigma})$, where $\boldsymbol{\mu}_P$, which quantifies the difference between P and Q, given by $\boldsymbol{\mu}_P = (\Delta_a, \Delta_a, \Delta_a, \Delta_a)'$; that is, the difference between the old form and new form populations is the same for all the components of the ability. The value Δ_a was set to .25, a borderline extreme ability difference that is rarely surpassed for large-scale

operational tests. From Figure 8, which is the pattern mostly observed in operational testing, this choice of $\boldsymbol{\mu}_P$ is reasonable.

2. Simulate scores on X in P, Y in Q, and those on A for both P and Q from the MIRT model using the draws of $\boldsymbol{\theta}$ from Step 1 and the generated item parameters for X, Y, and A.

   a. Perform equating of subscores and weighted averages using the previously mentioned methods using the scores of X in P, Y in Q and A in P and Q. To simplify matters, equating of weighted averages was computed only for the possible values of the corresponding subscore (i.e., for a simulation case with 12-item subtests, the equating was computed for 0, 1, . . ., 12).

   b. Ignore the scores on the items on the anchor test that belong to the third and fourth subscores, equate the subscores using Method 2 for subscore equating, and equate the weighted averages using Method 2 for equating weighted averages only using the items on the anchor test that belong to Subscores 1 and 2. This represents the nonrepresentative anchor test case that was considered in the operational data examples.

Computation of the true–population equating function. We can understand the true–population equating function for any subscore for a simulation case as the population value of the IRT observed score equating (e.g., Kolen & Brennan, 2004) for the subscore using linear interpolation as the continuization method. Consider a subscore X_s on Pseudotest X and the corresponding subscore Y_s on Pseudotest Y. The true–population equating function equating X_s to Y_s is the single-group equipercentile equating using the *true* raw subscore distributions corresponding to X_s and Y_s in a synthetic population T that places equal weights on P and Q. An iterative approach (Lord & Wingersky, 1984) was used to arrive at $P(X_s = x_s|\boldsymbol{\theta})$, which is an examinee's probability of obtaining a raw subscore of $X_s = x_s$ with ability $\boldsymbol{\theta}$. This approach involves the values of the item parameters; in this study, we used the generating item parameters for the items contributing to X_s. Once

$P(X_s = x_s|\boldsymbol{\theta})$ is computed, $r(x_s)$, we can assign a probability of a raw subscore of x_s to test X_s in population T, which can be obtained, using numerical integration, with the following:

$$r(x_s) = \int_{\boldsymbol{\theta}} P(X_s = x_s|\boldsymbol{\theta}) g_T(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2}$$

where $g_T(\boldsymbol{\theta}) = 0.5g_P(\boldsymbol{\theta}) + 0.5g_Q(\boldsymbol{\theta})$. As we assumed a simple structure, there is a function $P(X_s = x_s|\boldsymbol{\theta})$ of the corresponding component of $\boldsymbol{\theta}$ for any subscore s; that is, θ_s and the preceding integration results in a one-dimensional integration, with one over the marginal (standard normal) distribution of θ_s. The same approach provided us with $q(y_s)$, the probability of a raw score of y_s on Pseudotest Y in population T. The true raw score distributions $r(x_s)$ and $q(y_s)$, both discrete distributions, are then continuized using linear interpolation (Kolen & Brennan, 2004). Let us denote the corresponding continuized cumulative distributions, respectively, as $R(x_s)$ and $Q(y_s)$. The true equating function for subscore s is then obtained as $Q^{-1}(R(x_s))$. The true equating function is the same for each replication and correlation level; however, it varies with subscore length. Also, the true equating function for the subscores is used as the true equating function for the weighted averages as the weighted average is an estimate of the true subscore (Haberman, 2008b), and the true equating function computed earlier equates the true subscore distributions: the distributions of X_s and Y_s.

We can compute the true equating function using simulation (e.g., as in Sinharay & Holland, 2007) by generating scores of a huge sample of examinees on Pseudotests X and Y using the true item parameters, computing subscores and weighted averages, and performing a single-group equating of subscores or weighted averages. The values of the true equating function obtained using simulation were essentially identical for the equating of subscores. The two methods of computing the true equating functions for weighted averages had some differences. As an example, the simulation-based true equating function leads to slightly better values of performance criteria for the weighted averages than what is reported here. Henceforth, we do not discuss the simulation-based true equating function in this report.

Computation of the performance criteria: Equating bias, standard deviation, and root mean squared error. After we obtained equating results from M replications, we computed bias (a measure of systematic error in equating) and standard deviation (a measure of random error in equating) as performance criteria. For simulation, let $\hat{e}_i(x_s)$ be the equating function in the ith replication, providing the transformation of a raw (sub)score point x_s on X to the raw (sub)score scale of Y. Suppose $e(x_s)$ denotes the value of the corresponding true equating function. The bias at (sub)score point x_s is then obtained as follows:

$$\text{Bias}(x_s) = \frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x_s) - e(x_s)] = \bar{\hat{e}}(x_s) - e(x_s),$$

where

$$\bar{\hat{e}}(x_s) = \frac{1}{M}\sum_{i=1}^{M}\hat{e}_i(x_s).$$

We then obtain the corresponding standard deviation as follows:

$$\text{SD}(x_s) = \left\{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x_s) - \bar{\hat{e}}(x_s)]^2\right\}^{\frac{1}{2}}.$$

The corresponding RMSE is computed as follows:

$$\text{RMSE}(x_s) = \left\{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x_s) - e(x_s)]^2\right\}^{\frac{1}{2}}.$$

It can then be shown that

$$[\text{RMSE}(x_s)]^2 = [\text{SD}(x_s)]^2 + [\text{Bias}(x_s)]^2\,;$$

that is, the RMSE combines information from both random and systematic errors. The overall performance of a method for a simulation case can be judged by the overall (or weighted average of) bias, $\sum_{x_s} r(x_s)\text{Bias}(x_s)$; the overall standard deviation, $\sqrt{\sum_{x_s} r(x_s)\text{SD}^2(x_s)}$; and the overall RMSE, $\sqrt{\sum_{x_s} r(x_s)\text{RMSE}^2(x_s)}$.

Simulation Results

We present some summary statistics in Table 9 for the several simulation conditions, whereas Table 10 shows the weighted average of bias, standard deviation, and RMSE for Methods 1 and 2 for equating of subscores. The table also shows the values of the performance measures for no equating. Also, Table 11 shows the weighted average of bias, standard deviation, and RMSE for the equating of weighted averages. Table 10 shows the weighted average of bias, standard deviation, and RMSE for the equating of subscores when the scores on the items on the anchor test that belonged to the third and fourth subscores are ignored before performing the equating, that is, when the anchor test is nonrepresentative of the total test. Table 13 shows the weighted average of bias, standard deviation, and RMSE for the equating of weighted averages when the anchor test is nonrepresentative of the total test.

The values for the no equating column in Tables 11–13 are the same as those for the no equating column in Table 10. In interpreting the standard deviation and RMSE in the preceding tables, one should note that as test length increases, the standard error of measurement also increases; hence standard deviation and RMSE are also expected to increase.

The tables show the following:

- In Table 9, the percentage of the weighted averages that have added value is larger than 50% for all but two simulation cases, whereas the percentage of subscores that have added value is less than 50% for all but two simulation cases. Hence these simulation cases are appropriate for discussing equating of weighted averages.

- The methods seem to perform accurately. The bias, standard deviation, and RMSE of all the methods are small; that is, they are less than the DTM. This is in agreement with the results observed previously for the operational data examples.

- Any method leads to a substantial improvement over no equating for all simulation cases. Table 10 shows that the RMSE for no equating is more than the DTM for all but one simulation case; therefore the use of any of the suggested methods will most often lead to a more fair reported score.

- As we move toward the bottom left of Table 10, we notice a tendency for Method 1 of

Table 9
Summary Statistics From the Simulation Study

		Correlation		
Subscore length	Quantity	.70	.80	.90
12	α	.82	.84	.85
	r	.42	.48	.55
	PRMSE$_s$	.61	.61	.61
	PRMSE$_x$	.64	.71	.79
	PRMSE$_{sx}$	.72	.75	.81
	% sub	25	22	0
	% wtd	100	98	32
20	α	.89	.90	.91
	r	.50	.58	.65
	PRMSE$_s$	.72	.72	.72
	PRMSE$_x$	.69	.76	.84
	PRMSE$_{sx}$	.79	.82	.86
	% sub	66	25	7
	% wtd	100	100	46
30	α	.92	.93	.93
	r	.55	.63	.71
	PRMSE$_s$	.79	.79	.79
	PRMSE$_x$	.71	.79	.86
	PRMSE$_{sx}$	.84	.86	.89
	% sub	100	32	25
	% wtd	100	100	62

Note. Here α denotes average reliability of the total score; r denotes 100× the average correlation between the subscores; % sub denotes the overall percentage of subscores that have added value; and % wtd denotes the overall percentage of weighted averages that have added value.

Table 10

Overall Bias, Standard Deviation, and Root Mean Squared Error (RMSE; All Multiplied by 100) for Equating of Subscores in the Simulation Study

			Correlation = .70				Correlation = .80				Correlation = .90			
Subscore			Subscore				Subscore				Subscore			
length	Measure	Method	1	2	3	4	1	2	3	4	1	2	3	4
12	Bias	1	–15	–9	–13	–7	–15	–10	–13	–7	–15	–10	–13	–7
		2	5	4	7	6	4	3	6	5	3	1	5	4
	SD	1	10	12	10	11	11	12	10	11	11	12	11	11
		2	10	11	10	10	10	11	10	10	10	11	10	10
	RMSE	1	19	16	16	15	18	16	17	15	18	16	17	16
		2	12	12	12	12	11	11	12	11	10	11	11	10
		None	65	57	63	47	65	57	63	47	65	57	63	47
20	Bias	1	–12	–6	–6	–9	–12	–6	–6	–9	–12	–7	–6	–10
		2	15	14	15	14	12	10	13	11	9	7	10	8
	SD	1	15	17	14	15	16	16	14	15	15	16	14	15
		2	15	19	15	14	14	18	14	14	13	16	14	13
	RMSE	1	20	18	15	18	20	18	16	18	20	17	16	18
		2	21	24	22	20	19	21	19	18	16	18	18	16
		None	89	76	84	80	89	76	84	80	89	76	84	80
30	Bias	1	–9	–9	–6	–8	–9	–9	–5	–7	–9	–9	–5	–7
		2	18	19	22	19	14	13	19	15	10	8	15	11
	SD	1	21	20	20	19	21	20	19	19	21	21	19	19
		2	22	21	19	19	20	19	18	18	19	18	19	17
	RMSE	1	23	23	21	22	23	23	21	22	23	23	20	23
		2	29	28	31	28	25	23	27	24	22	20	25	21
		None	126	123	136	138	126	123	136	138	126	123	136	138

Note. None refers to no equating. For these cases, SD = 0 and bias = RMSE, and hence bias and SD are not included in the table.

Table 11

Overall Bias, Standard Deviation, and RMSE (All Multiplied by 100) for Equating of Weighted Averages in the Simulation Study

			Correlation = .70				Correlation = .80				Correlation = .90			
			Subscore				Subscore				Subscore			
SL	Mea.	Meth.	1	2	3	4	1	2	3	4	1	2	3	4
12	Bias	1	−30	−24	−27	−14	−28	−23	−26	−14	−26	−22	−26	−15
		2	−19	−18	−19	−16	−17	−17	−19	−15	−16	−17	−18	−16
		3	−21	−21	−22	−18	−18	−19	−20	−15	−16	−18	−18	−16
	SD	1	13	17	17	12	13	13	16	12	13	13	16	12
		2	15	16	17	15	14	15	16	15	14	14	15	14
		3	15	15	17	15	14	15	16	15	14	15	15	14
	RMSE	1	36	33	39	41	34	32	37	38	32	30	36	37
		2	25	25	31	25	23	24	30	24	22	23	29	24
		3	27	27	34	26	25	25	31	24	23	24	30	24
20	Bias	1	−30	−23	−25	−27	−28	−22	−25	−27	−26	−22	−25	−25
		2	−12	−10	−14	−11	−11	−12	−15	−12	−11	−14	−16	−12
		3	−17	−18	−18	−15	−14	−16	−17	−14	−11	−15	−16	−12
	SD	1	20	19	17	18	18	17	18	19	18	18	27	18
		2	20	22	19	18	17	20	18	17	17	18	26	16
		3	19	19	19	17	17	18	18	17	16	17	26	15
	RMSE	1	40	36	38	35	36	35	37	34	34	34	42	33
		2	25	26	28	24	22	26	27	24	21	25	34	24
		3	27	28	29	24	23	26	28	24	21	25	34	22
30	Bias	1	−32	−30	−33	−31	−30	−29	−31	−30	−27	−28	−30	−28
		2	−12	−08	−13	−10	−13	−12	−14	−12	−13	−14	−15	−13
		3	−20	−20	−21	−18	−17	−19	−18	−16	−13	−17	−17	−14
	SD	1	23	21	24	20	23	22	22	21	22	22	22	20
		2	26	23	23	22	23	22	21	20	20	21	20	19
		3	23	21	24	21	21	21	21	20	20	21	20	18
	RMSE	1	45	40	46	42	42	41	42	41	39	40	41	41
		2	32	27	29	27	28	29	27	27	25	30	27	31
		3	32	31	33	30	29	31	29	28	25	30	28	29

Note. The RMSE for no equating for any simulation case shown in the table is identical to the RMSE for the same simulation case in Table 10. Mea. = measure, meth. = method, SL = subscore length.

Table 12
Overall Bias, Standard Deviation, and RMSE (All Multiplied by 100) for Equating of Subscores When the Anchor Is Nonrepresentative in the Simulation Study

			Correlation = .70				Correlation = .80				Correlation = .90			
Subscore			Subscore				Subscore				Subscore			
length	Measure	Method	1	2	3	4	1	2	3	4	1	2	3	4
12	Bias	2	−2	−4	−0	−1	−3	−5	−1	−2	−3	−6	−1	−2
	SD	2	11	12	11	12	11	12	11	12	10	11	10	10
	RMSE	2	11	12	11	12	11	13	11	12	11	13	10	11
20	Bias	2	5	3	6	3	4	1	5	2	3	–1	4	1
	SD	2	16	18	17	16	15	17	16	16	15	17	16	15
	RMSE	2	17	18	18	17	16	17	17	16	15	17	16	15
30	Bias	2	9	7	14	10	6	4	11	8	4	1	9	5
	SD	2	22	19	23	24	21	20	23	23	20	18	22	20
	RMSE	2	24	21	28	27	23	21	26	25	21	19	24	22

Table 13
Overall Bias, Standard Deviation, and RMSE (All Multiplied by 100) for Equating of Weighted Averages When the Anchor Is Nonrepresentative in the Simulation Study

			Correlation = .70				Correlation = .80				Correlation = .90			
			Subscore				Subscore				Subscore			
SL	Mea.	Meth.	1	2	3	4	1	2	3	4	1	2	3	4
12	Bias	2	−23	−23	−23	−21	−21	−22	−22	−19	−20	−22	−21	−20
	SD	2	15	16	18	16	14	16	17	16	14	14	15	14
	RMSE	2	29	29	34	29	27	28	33	28	25	27	31	27
20	Bias	2	−19	−19	−19	−19	−17	−19	−19	−18	−15	−20	−19	−17
	SD	2	20	21	19	19	17	19	18	18	17	19	27	17
	RMSE	2	29	30	31	28	26	29	30	28	24	29	36	27
30	Bias	2	−19	−17	−20	−18	−19	−19	−19	−18	−17	−20	−20	−18
	SD	2	26	22	27	26	24	22	25	25	23	22	23	22
	RMSE	2	36	30	35	34	33	32	33	33	30	33	32	35

Note. Mea. = measure, meth. = method, SL = subscore length.

equating of subscores to perform better when compared to Method 2. This means that when subscores are longer and more uncorrelated to each other, a subscore is more likely to stand on its own, and an equating with the anchor items contributing only to that subscore is expected to be more accurate than an equating with the total (scaled) score as an anchor score.

- Both Methods 2 and 3 for equating weighted averages perform better than Method 1, as is clear from Table 11. Among Methods 2 and 3, the former performs slightly better than the latter overall—the RMSE is more often smaller for the former than for the latter.

- As the correlation level increases, the RMSE for Method 2 for equating of subscores becomes less because the total score becomes more similar to a subscore as correlation among the subscores increases. The same phenomenon is observed for all the methods for equating of weighted averages but not for Method 1 of equating of subscores.

- If Tables 10 and 11 are compared, the performance measures for any simulation case have higher absolute values for weighted averages than for subscores. Even so, the values of the measures for the weighted averages are less than the DTM.

- Bias, standard deviation, and RMSE are larger for longer subtests than for shorter subtests.

- Tables 10 and 13 show that the equating appears to be accurate even for the simulation cases with a nonrepresentative anchor test. The quality of equating with nonrepresentative anchors is mostly worse than that with representative anchors; this is clear from a comparison of Tables 11 and 13. However, equating with nonrepresentative anchors is much better than no equating at all. This is clear from a comparison of the values given in Tables 12 and 13 with the values for no equating shown in Table 10. The small error of equating with nonrepresentative anchors in the simulations is in agreement with results from the operational data sets because, we believe, the anchor test is supposed to reflect the differences in the two populations on the subscores to be equated, and the assumption of equal difference for all subscores in the preceding simulation causes the scaled total score, which is used as the anchor score in Method 2, to reflect the difference accurately.

Additional Simulations

The preceding simulations assumed that $\boldsymbol{\mu}_P = (\Delta_a, \Delta_a, \Delta_a, \Delta_a)'$; that is, the difference in the ability of the two populations is the same for all subscores. The simulations also assumed that the difference in the difficulty of the two tests is the same for all four content areas. In the literature, Sinharay and Holland (2007) discussed other possible patterns of differences between the two populations and the two tests when data are generated from a MIRT model, and for our purposes here, it is of interest to determine whether the results reported in Tables 10–13 hold under those patterns.

Table 14 considers the case when the two populations are of the same ability, that is, all the components of $\boldsymbol{\mu}_P$ are 0 and the two tests are of the same difficulty for all four content areas. This is the ideal equating scenario because there is no need to equate in this case. Table 14 shows the values of RMSE for Method 2 for equating subscores and Method 2 for equating weighted averages, both for the case when the anchor test is representative and for the case when the anchor test is nonrepresentative. In this case, the methods' comparative performance is similar to what is shown in Tables 10–13, and hence results for other methods are not provided in this report. Table 14 shows the results for the three lengths of subscores (12, 20, and 20) and the two levels of correlations (70 and 90). These values are mostly slightly lower than those reported in Tables 10–13.

Next we performed some additional simulations under different patterns. Different patterns of difference in difficulty of the two tests did not affect results, so the pattern was set as in preceding simulations; that is, Pseudotest X is more difficult than Pseudotest Y in all four content areas. However, different patterns of differences in ability of the two populations substantially affected the results; therefore we report results for two of these patterns. Table 15 shows results for the case in which $\boldsymbol{\mu}_P = (0.1, 0.15, 0.2, 0.25)$, which is likely to occur in practice, as demonstrated by Figure 8. Table 16 provides results for the case in which $\boldsymbol{\mu}_P = (0.1, 0.1, -0.1, -0.1)$. This kind of pattern was discussed by Klein and Jarjoura (1985; see their Figures 2 and 3). The preceding results were obtained when the anchor test length was 40% of the total test length. Table 17 provides results for the case in which the anchor test length is 24 (i.e., six items per subtest) and $\boldsymbol{\mu}_P = (0.25, 0.25, 0.25, 0.25)$. The table helps in understanding how the methods are performed as the total test length to anchor test length decreases from 50% (in the case of 12 items per subscore) to 20% (in the case of 30 items per subscore).

Table 14
Overall RMSE (Multiplied by 100) for Equating of Subscores (Method 2) and Weighted Averages (Method 2) When the Two Populations Are of the Same Ability and the Two Tests Are of Equal Difficulty

			Correlation = .70				Correlation = .90			
Subscore	Subscore	Anchor	Subscore				Subscore			
length	or wtd. av.	type	1	2	3	4	1	2	3	4
12	Subscore	Representative	12	12	11	12	11	12	11	11
12	Wtd. av.	Representative	18	17	25	18	16	14	23	18
12	Subscore	Nonrepresentative	12	13	12	13	12	12	11	11
12	Wtd. av.	Nonrepresentative	18	17	25	19	16	15	23	18
12	Both	None	20	10	09	03	20	10	09	03
20	Subscore	Representative	15	18	15	15	14	17	14	13
20	Wtd. av.	Representative	21	20	23	19	20	17	23	18
20	Subscore	Nonrepresentative	16	18	17	17	15	17	16	15
20	Wtd. av.	Nonrepresentative	22	20	24	20	21	18	23	19
20	Both	None	20	11	04	08	20	11	04	08
30	Subscore	Representative	22	22	20	20	19	19	19	17
30	Wtd. av.	Representative	26	23	24	23	22	21	22	22
30	Subscore	Nonrepresentative	23	21	25	25	21	19	22	21
30	Wtd. av.	Nonrepresentative	27	22	28	27	24	21	25	24
30	Both	None	20	07	17	20	20	07	17	20

Note. RMSE = root mean squared error, wtd. av. = weighted average.

Table 15

Overall RMSE (Multiplied by 100) for Equating of Subscores and Weighted Averages When $\boldsymbol{\mu}_P = (0.1, 0.15, 0.2, 0.25)$

	Subscore		Correlation = .70				Correlation = .90			
Subscore	or wtd. av.	Anchor	Subscore				Subscore			
length	and method	type	1	2	3	4	1	2	3	4
12	Subscore 1	Representative	12	13	14	15	12	13	15	16
12	Subscore 2	Representative	20	15	10	14	18	13	10	15
12	Wtd. av. 1	Representative	26	28	36	41	24	24	34	36
12	Wtd. av. 2	Representative	19	19	31	31	16	17	30	31
12	Wtd. av. 3	Representative	20	21	32	31	16	18	31	31
12	Subscore 2	Nonrepresentative	11	14	16	26	11	14	15	26
12	Wtd. av. 2	Nonrepresentative	21	24	36	39	18	23	34	38
12	Both	None	66	57	64	47	66	57	64	47
20	Subscore 1	Representative	16	17	15	18	16	16	15	18
20	Subscore 2	Representative	35	27	16	20	31	22	14	21
20	Wtd. av. 1	Representative	28	31	36	35	26	28	34	36
20	Wtd. av. 2	Representative	24	23	30	37	24	19	29	39
20	Wtd. av. 3	Representative	22	22	30	33	22	19	29	37
20	Subscore 2	Nonrepresentative	19	19	23	39	18	18	23	39
20	Wtd. av. 2	Nonrepresentative	22	28	38	52	20	27	37	52
20	Both	None	91	77	85	80	91	77	85	80
30	Subscore 1	Representative	21	21	21	21	21	20	20	23
30	Subscore 2	Representative	49	35	21	27	42	27	19	31
30	Wtd. av. 1	Representative	34	31	42	43	30	32	37	45
30	Wtd. av. 2	Representative	34	26	32	49	29	24	30	54
30	Wtd. av. 3	Representative	24	24	34	41	25	23	30	50
30	Subscore 2	Nonrepresentative	27	22	32	55	24	23	32	57
30	Wtd. av. 2	Nonrepresentative	27	31	49	72	24	33	46	73
30	Both	None	129	125	137	138	129	125	141	141

Note. RMSE = root mean squared error, wtd. av. = weighted average.

Table 16
Overall RMSE (Multiplied by 100) for Equating of Subscores and Weighted Averages When $\boldsymbol{\mu}_P = (0.1, 0.1, -0.1, -0.1)$

	Subscore		Correlation = .70				Correlation = .90			
Subscore	or wtd. av.	Anchor	Subscore				Subscore			
length	and method	type	1	2	3	4	1	2	3	4
12	Subscore 1	Representative	12	13	11	11	12	13	11	11
12	Subscore 2	Representative	20	25	17	20	20	25	17	19
12	Wtd. av. 1	Representative	26	27	30	45	23	22	27	40
12	Wtd. av. 2	Representative	29	31	27	23	28	29	26	23
12	Wtd. av. 3	Representative	28	26	28	21	27	25	26	22
12	Subscore 2	Nonrepresentative	11	12	32	34	11	12	31	33
12	Wtd. av. 2	Nonrepresentative	23	23	32	30	20	20	31	31
12	Both	None	66	57	65	49	66	57	65	49
20	Subscore 1	Representative	16	17	15	16	16	16	15	15
20	Subscore 2	Representative	31	39	30	34	31	39	30	34
20	Wtd. av. 1	Representative	29	29	32	23	26	27	30	21
20	Wtd. av. 2	Representative	39	44	32	33	36	43	32	34
20	Wtd. av. 3	Representative	31	29	26	25	33	31	29	32
20	Subscore 2	Nonrepresentative	16	18	56	64	15	17	55	63
20	Wtd. av. 2	Nonrepresentative	25	24	46	54	21	23	48	56
20	Both	None	91	77	87	82	91	77	87	82
30	Subscore 1	Representative	21	20	19	18	21	20	18	19
30	Subscore 2	Representative	50	59	45	50	48	58	44	49
30	Wtd. av. 1	Representative	35	29	31	28	33	31	27	26
30	Wtd. av. 2	Representative	57	62	41	46	56	61	40	47
30	Wtd. av. 3	Representative	39	32	29	27	48	39	32	42
30	Subscore 2	Nonrepresentative	23	21	88	97	20	20	85	94
30	Wtd. av. 2	Nonrepresentative	31	25	70	64	29	27	71	85
30	Both	None	129	125	141	141	129	125	141	141

Note. RMSE = root mean squared error, wtd. av. = weighted average.

Table 17
Overall RMSE (Multiplied by 100) for Equating of Subscores and Weighted Averages When the Anchor Test Length Is 24 and $\mu_P = (0.25, 0.25, 0.25, 0.25)$

	Subscore		Correlation = .70				Correlation = .90			
Subscore	or wtd. av.	Anchor	Subscore				Subscore			
length	and method	type	1	2	3	4	1	2	3	4
12	Subscore 1	Representative	14	15	15	12	14	15	15	11
12	Subscore 2	Representative	14	14	15	12	12	12	13	11
12	Wtd. av. 1	Representative	32	35	36	28	28	33	34	27
12	Wtd. av. 2	Representative	24	26	30	24	20	25	29	25
12	Wtd. av. 3	Representative	25	28	32	24	20	25	29	24
12	Subscore 2	Nonrepresentative	11	12	12	10	10	12	11	10
12	Wtd. av. 2	Nonrepresentative	26	30	32	27	23	28	30	26
12	Both	None	65	45	60	48	65	45	60	48
20	Subscore 1	Representative	25	27	21	23	25	27	22	22
20	Subscore 2	Representative	19	20	21	18	15	17	17	15
20	Wtd. av. 1	Representative	45	47	43	43	40	45	41	39
20	Wtd. av. 2	Representative	28	29	29	27	24	28	28	25
20	Wtd. av. 3	Representative	32	36	34	30	24	28	28	25
20	Subscore 2	Nonrepresentative	17	19	18	17	17	19	16	16
20	Wtd. av. 2	Nonrepresentative	34	36	35	34	30	35	33	31
20	Both	None	89	85	82	81	89	85	82	81
30	Subscore 1	Representative	33	38	41	38	33	37	41	39
30	Subscore 2	Representative	24	27	23	25	20	24	20	20
30	Wtd. av. 1	Representative	58	57	65	59	53	57	60	55
30	Wtd. av. 2	Representative	33	36	37	37	29	39	35	33
30	Wtd. av. 3	Representative	40	45	47	41	30	41	38	31
30	Subscore 2	Nonrepresentative	24	29	26	28	24	29	25	26
30	Wtd. av. 2	Nonrepresentative	41	43	46	46	37	46	42	40
30	Both	None	126	123	138	123	126	123	138	123

Note. RMSE = root mean squared error, wtd. av. = weighted average.

Tables 15 and 16, which have similar results, draw a slightly different picture concerning the performance of the suggested equating methods from that drawn by Tables 10–13 in that the former show the following:

1. The RMSEs are often much larger in Tables 15 and 16 in comparison to those in Tables 10–13. For example, the largest RMSE in Tables 10 and 12 for any equating method was .31, whereas for subscore equating, the RMSE can be as large as .97 in Table 16. Even then, equating with some method is better than no equating, according to Tables 15 and 16. Also, the RMSE is less than the DTM of 0.50 more than 90% of the time.

2. Unlike Table 10, when the content is representative, Method 2 for subscore equating

is worse than Method 1 for almost all cases in Tables 15 and 16, including for 12-item subtests. This is expected. In the simulations underlying Tables 15 and 16, the difference in the total scaled score between the two populations does not reflect the difference in any single subscore between the two populations. For example, when $\boldsymbol{\mu}_P = (0.1, 0.1, -0.1, -0.1)$, which led to Table 16, the average total scores for the two populations are close, whereas the average of any of the subscores differs over the two populations. Hence, in Tables 15 and 16, the total score does not perform as an anchor as well as it does in Tables 10–13.

3. Unlike Table 11, Method 2 for equating weighted averages is often the worst of the three methods, especially as shown in Table 16.

4. Unlike in Table 12, a nonrepresentative anchor often leads to much worse equating of subscores compared to the representative anchor of Tables 15 and 16. For example, the RMSE for the fourth 30-item subscore for correlation = .70 in Table 16 is .97 for the nonrepresentative anchor case compared to .18 and .50 (Methods 1 and 2, respectively) for the representative anchor case. This is because the total scaled score, which acts as the anchor score, in the nonrepresentative anchor case is based on the first two subscores, and P is better than Q in these two subscores (the first two components of $\boldsymbol{\mu}_P$ are positive); however, P is worse than Q on the third and fourth subscores, so a substantial equating error is expected in the equating of these subscores. The equating error for the first two subscores is small for the nonrepresentative anchor case because the total score accurately reflects the differences between the two populations for these two subscores.

5. A comparison of results from Tables 17 and 10 for subtest lengths 20 and 30 supports the findings of Puhan and Liang (in press), which suggest that Method 2 of subscore equating performs better than Method 1 when the proportion of a subtest that is common between the old and new forms is small. In Table 10, for subtest length 30, for which the anchor test length is 48, Method 1 performs better than Method 2. However, in Table 17, for subtest length 30, for which the anchor test length is 24,

Method 1 performs worse than Method 2.

Conclusions

This report reviewed several methods, working under the NEAT design, for the equating of subscores and suggested several methods for the equating of weighted averages. The report also examined the performance of the discussed and described methods using several operational and simulated data sets. The results demonstrated that the suggested methods perform quite accurately; that is, the extent of equating error (both systematic and random) is small in most situations. Even in the case when the anchor test is nonrepresentative of the test (i.e., some content areas covered in the test were not covered in the anchor test), the suggested methods still perform quite well in most cases. Through the study, we also demonstrated that borrowing information from other subscores in an anchor score can lead to an increase in accuracy of the equating of subscores and weighted averages.

The results of this report are subject to the usual limitations of simulation studies; however, the following facts make our simulations somewhat realistic:

- In essence, the results from the simulation study are shown to be in agreement with results from the operational data sets. As an example, when the difference in the populations is in the same direction for all the subscores, the equating error is small both for the simulated data and for the operational data.

- To make the data sets more realistic, the simulations used item parameters estimated from operational data to generate the simulated data sets.

- The data sets simulated in our study were found to reproduce adequately the raw subscore distributions of the operational data set. Because the functions equating subscores or weighted averages are completely determined by the corresponding raw subscore distributions, our simulations are realistic for our purpose.

- Haberman, von Davier, and Lee (2008) found that MIRT models fit operational data better than univariate IRT models and that they provide a reasonably good fit to operational data sets; therefore the data simulated from a MIRT model in this study can be expected to retain the important features of the operational data reasonably well.

Several issues should be examined further:

- More operational data sets and simulated data sets, especially those different in nature from those considered in this report, should be analyzed using the methods we suggest here. For example, data involving polytomous items and internal anchor tests were not analyzed in this report and should be studied further.
- Also to be examined is population invariance of equating (e.g., Dorans & Holland, 2000) of subscores and weighted averages. Practitioners often argue that differences between subscores vary over different subgroups, in which case, the equating method that uses the total score as an anchor score may be more prone to lack of invariance.
- This report considered chain equipercentile equating using linear interpolation (Kolen & Brennan, 2004). Other equating methods, for example, kernel equating (von Davier, Holland, & Thayer, 2003) and equating using continuous exponential families (Haberman, 2008a), could also be applied—these methods would replace the linear interpolation (Kolen & Brennan, 2004) that was used in this report by more elegant smoothing techniques. The existing software for equating using continuous exponential families (Haberman, 2008a) has the advantage that it can handle fractional scores.
- It would be useful to perform a study of scale drift in the context of equating of subscores and weighted averages. One method considered here may be more prone to scale drift than another method.
- A study of whether the weights on the weighted averages vary too much over the different forms of a test may be worthwhile. If they do not, then it may be possible to fix the weights once and then use those weights in the future to report weighted averages. In addition, the equating performance of weighted averages with fixed weight should be studied.

References

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Rep. No. 94-10). Princeton, NJ: ETS.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). *A comparison of subscale score augmentation methods using empirical data.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Fransisco, CA.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item types. *Journal of Educational Measurement*, *35*, 137–154.

ETS. (2007). *General science: Content essays (0433), test analysis form 4CPX1* (ETS Research Rep. No. SR-2007-109). Princeton, NJ: ETS.

Haberman, S. J. (2008a). *Continuous exponential families: An equating tool* (ETS Research Rep. No. RR-08-05). Princeton, NJ: ETS.

Haberman, S. J. (2008b). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the neat design. *Journal of Educational Measurement*, *45*, 17–43.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement*, *22*, 197–206.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Livingston, S. A. (1994). *Equating constructed-response tests through a multiple-choice anchor: A small-scale empirical study* (ETS Research Rep. No. SR-94-100). Princeton, NJ: ETS.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, *8*, 453–461.

Puhan, G., & Liang, L. (in press). *Equating subscores under the non equivalent anchor test (NEAT) design* (ETS Research Rep.). Princeton, NJ: ETS.

Puhan, G., Moses, T. P., Grant, M. C., & McHale, F. (2009). Small-sample equating using a single-group nearly equivalent test (signet) design. *Journal of Educational Measurement*, *46*, 344–362.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Amsterdam: North-Holland.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (ETS Research Rep. No. RM-08-18). Princeton, NJ: ETS.

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*, 249–275.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data* (ETS Research Rep. No. RR-06-02). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). *The kernel method of test* *equating.* New York: Springer.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (2001). Augmented scores—“borrowing strength” to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum Associates.

Notes

[1] Of course, one must check the accuracy of such equating before the implementation of the methods for such situations.

[2] Of course, the accuracy of such equating has to be checked before its implementation.

[3] This is important because several testing programs operationally report subscores that are not equated.

[4] If weighted averages are operationally used, they are likely to be rounded to their nearest integers; therefore this strategy is reasonable.

[5] When the level of correlation is .90, the (3,4)th element of $\mathcal{C}$ was changed to .85 before this calculation so that Σ was ensured to be positive definite.

Appendix

Description of the Methodology of Haberman (2008b)

In this appendix, we describe the methodology of Haberman (2008b) and Haberman, Sinharay, and Puhan (2009), used in this report to determine whether and how to report examinee-level subscores. The analysis involves the observed subscore s, the true subscore s_t, the observed total score x, and the true total score x_t. It is assumed that s_t, x_t, $s - s_t$, and $x - x_t$ all have positive variances. As usual in CTT, s and s_t have common mean $E(s)$, x and x_t have common mean $E(x)$, and the true scores s_t and x_t are uncorrelated with the errors $s - s_t$ and $x - x_t$. Let $\rho(a, b)$ denote the correlation between a and b. It is assumed that the true subscore s_t and true total score x_t are not collinear so that $|\rho(s_t, x_t)|$ is less than 1. This assumption also implies that $|\rho(s, x)| < 1$. Haberman (2008b) considered several approaches for estimation of the true score s_t.

In the first approach, s_t is estimated by the constant $E(s)$ so that the corresponding mean squared error in estimation is $E[s_t - E(s)]^2 = \sigma^2(s_t)$. In the second, the linear regression

$$s_s = E(s) + \rho^2(s_t, s)[s - E(s)]$$

of s_t on the observed subscore s estimates s_t, and the corresponding mean squared error is $E(s_t - s_s)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, s)]$, where $\rho^2(s_t, s)$ is the reliability of the subscore. In the third approach, the linear regression

$$s_x = E(s) + \rho(s_t, x)[\sigma(s_t)/\sigma(x)][x - E(x)]$$

of s_t on the observed total score x estimates s_t, and the corresponding mean squared error is $E(s_t - s_x)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, x)]$.

Haberman (2008b) compared the last two approaches with respect to their PRMSE. Relative to using $E(s)$, the PRMSE corresponding to the use of s_s as the estimate of s_t is given by

$$\text{PRMSE}_s = \frac{\sigma^2(s_t) - \sigma^2(s_t)[1 - \rho^2(s_t, s)]}{\sigma^2(s_t)} = \rho^2(s_t, s),$$

which is the reliability of the subscore. Relative to using $E(s)$, the PRMSE corresponding to the use of s_x as the estimate of s_t is PRMSE$_t = \rho^2(s_t, x)$, where

$$\rho^2(s_t, x) = \rho^2(s_t, x_t)\rho^2(x_t, x), \tag{A1}$$

where $\rho^2(x_t, x)$ is the total score reliability. The computation of $\rho^2(s_t, x_t)$ is described shortly.

Haberman (2008b) argued on the basis of these results that the true subscore is better approximated by s_x (which is an estimate based on the total score) than by s_s (which is an estimate based on the subscore) if $\rho^2(s_t, s)$ is smaller than $\rho^2(s_t, x)$, and hence subscores should not be reported in that case.

The fourth approach consists of reporting an estimate of the true subscore s_t based on the linear regression s_{sx} of s_t on both the observed subscore s and the observed total score x. The regression is given by

$$s_{sx} = E(s) + \beta[s - E(s)] + \gamma[x - E(x)],$$

where

$$\gamma = \frac{\sigma(s)}{\sigma(x)}\rho(s_t, s)\tau,$$

$$\tau = \frac{\rho(x_t, x)\rho(s_t, x_t) - \rho(s, x)\rho(s_t, s)}{1 - \rho^2(s, x)},$$

$$\beta = \rho(s_t, s)[\rho(s_t, s) - \rho(s, x)\tau].$$

The mean squared error is then $E(s_t - s_{sx})^2 = \sigma^2(s_t)\{1 - \rho^2(s_t, s) - \tau^2[1 - \rho^2(s, x)]\}$ so that the PRMSE relative to $E(s)$ is given by

$$\text{PRMSE}_{st} = \rho^2(s_t, s_{sx}) = \rho^2(s_t, s) + \tau^2[1 - \rho^2(s, x)].$$

Computation of $\boldsymbol{\rho^2(s_t, x_t)}$

The quantity $\rho^2(s_t, x_t)$ can be expressed as

$$\rho^2(s_t, x_t) = \frac{[\text{Cov}(s_t, x_t)]^2}{V(s_t)V(x_t)}.$$

The variances are computed by multiplying the observed variance by the reliabilities; for example,

$$V(s_t) = \rho^2(s_t, s) \times \text{observed variance of } s.$$

The covariance $\mathrm{Cov}(s_t, x_t)$ can be expressed, where s_{kt} denotes the true kth subscore, as

$$\mathrm{Cov}(s_t, x_t) = \mathrm{Cov}\left(s_t, \sum_k s_{kt}\right) = \sum_k \mathrm{Cov}\left(s_t, s_{kt}\right).$$

The right-hand side of the equation is the sum of the tth row of C_T, the covariance matrix between the true subscores. The off-diagonal elements of C_T are the same as those of the covariance matrix between the observed subscores; the kth diagonal element of C_T is obtained as variance of the kth observed subscore × reliability of the kth subscore.