JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Summer 2011, Vol. 48, No. 2, pp. 122–145

Equating of Augmented Subscores

Sandip Sinharay and Shelby J. Haberman
Educational Testing Service

Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. Haberman (2008b) suggested reporting an augmented subscore that is a linear combination of a subscore and the total score. Sinharay and Haberman (2008) and Sinharay (2010) showed that augmented subscores often lead to more accurate diagnostic information than subscores. In order to report augmented subscores operationally, they should be comparable across the different forms of a test. One way to achieve comparability is to equate them. We suggest several methods for equating augmented subscores. Results from several operational and simulated data sets show that the error in the equating of augmented subscores appears to be small in most practical situations.

There is an increasing interest in subscores, which are scores on subtests or subareas, because of their potential diagnostic value. The U.S. Government's No Child Left Behind Act of 2001 demands, among other things, that students should receive diagnostic reports that allow teachers to address their specific academic needs; subscores could potentially be used in such a diagnostic report.

Haberman (2008b) suggested a method based on classical test theory to determine if subscores have added value. Haberman (2008b) also suggested reporting a linear combination of a subscore and the total score instead of reporting the subscore. The linear combinations suggested by Haberman (2008b) are special cases of the augmented subscores of Wainer, Sheehan, and Wang (2000) and will be referred to as *augmented subscore*s henceforth.[1] Augmented subscores may be difficult to explain and clients may not like the idea that an examinee's reported Reading (augmented sub)score is based not only on the examinee's actual Reading subscore, but also on the examinee's actual Listening subscore. However, Sinharay and Haberman (2008), Sinharay (2010), and Puhan, Sinharay, Haberman, and Larkin (2010), who studied several operational and simulated data sets, found that augmented subscores often lead to more accurate diagnostic information than subscores.

In order for augmented subscores to be reported operationally, they should be comparable across the different forms of a test. One way to achieve comparability is to equate them. There is a lack of research on equating of augmented subscores. For a single-group design or equivalent-groups design, the equating is straightforward and will involve, for example, a direct linear equating or equipercentile equating. For the nonequivalent groups with anchor test design, equating of augmented subscores is not straightforward, and approaches that are used to equate the total score may not be appropriate for equating augmented subscores. First, the anchor test usually includes only a few items on any single subarea. For example, consider a test whose total score, which is the sum of four subscores, is equated using a 20-item anchor test. The anchor test would contain about five items that belong to each subarea. Suppose that the interest is in equating the first augmented subscore. Five items belonging to the first subarea are too few to lead to accurate anchor test equating of the first augmented subscore (according to the recommendations of, e.g., Kolen and Brennan, 2004, p. 271). Second, augmented subscores have fractional values whereas most equating software packages such as CIPE (Kolen and Brennan, 2004) and LOGLIN/KE (Chen, Lu, Thayer, & Han, 2007) can handle only integer-valued scores.[2] So it is not straightforward to use standard software packages for equating augmented subscores. Thus, it is not clear how one can equate augmented subscores using existing methods and software packages. The primary goal of this paper is to examine this issue of equating of augmented subscores.

There are several operational tests that report subscores (or section scores) and use a nonequivalent groups with anchor test design, but the anchor test does not cover one or two of the subareas, that is, the anchor test is not representative of the total test. For example, four section scores are reported for TOEFL®: Reading, Listening, Speaking, and Writing. Because of concerns about item exposure, a TOEFL form usually includes some Reading and Listening items common with a previous form, but no Speaking and Writing items common with any previous form. As a result, the TOEFL Reading and Listening scores are equated using the common items, but there is no way to equate the TOEFL Speaking and Writing scores. One goal of this paper is to examine the extent of the error in equating (both systematic error and random error) the augmented subscores from such a test, especially those corresponding to the subareas that are not represented in the anchor test.

The next section discusses the classical-test-theory-based approach of Haberman (2008b) and the augmented subscores. The Methods section discusses our suggested methods for equating of augmented subscores. The Application section discusses results from the application of the suggested methods to an operational data set. The Simulation Study section discusses results of the application of the suggested methods to several simulated data sets. The last section discusses the conclusions and recommendations.

The Classical-test-theory-based Approach and Augmented Subscores

Let us denote the subscore and the total score of an examinee as s and x, respectively. Let x_t denote the corresponding true total score. Haberman (2008b), taking a classical test theory viewpoint, 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:

- The estimate $s_s = \bar{s} + \alpha(s - \bar{s})$ that is based on the observed subscore, where $\bar{s}$ is the average subscore for the sample of examinees and α is the reliability of the subscore.
- The estimate $s_x = \bar{s} + \frac{\sigma(s_t)}{\sigma(x)}\rho(x_t, x)\rho(s_t, x_t)(x - \bar{x})$ that is based on the observed total score, where $\sigma(u)$ and $\rho(u, v)$, respectively, denote the standard deviation of u and the correlation between u and v.

- The augmented subscore $s_{sx} = \bar{s} + a(x - \bar{x}) + b(s - \bar{s})$ that is based on both the subscore and the total score, where

$$a = \frac{\sigma(s)}{\sigma(x)}\rho(s_t, s)\tau,$$
$$b = \rho(s_t, s)[\rho(s_t, s) - \rho(s, x)\tau],$$

and

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

The average of the augmented subscore for a sample is the same as the average of the subscores. So the augmented subscores are of the same magnitude as the subscores. However, the variance of the augmented subscores is smaller than that of the subscores; that is because the augmented subscore corresponding to an extremely low or high subscore is pooled to the mean and not as extreme.

To compare the performances of s_s, s_x, and s_{sx} as estimates of s_t, Haberman (2008b) suggested the use of the proportional reduction in mean squared error (PRMSE). Let us denote the PRMSEs corresponding to s_s, s_x, and s_{sx} as $PRMSE_s$, $PRMSE_x$, and $PRMSE_{sx}$, respectively. The PRMSE is conceptually similar to reliability. The quantity $PRMSE_s$ is identical to the subscore reliability. The larger the PRMSE, the more accurate is the estimate. Haberman (2008b) recommended that for a subscore to have *added value*, $PRMSE_s$ has to be larger than $PRMSE_x$. He also recommended that for the augmented subscore to have added value, $PRMSE_{sx}$ has to be substantially larger than both $PRMSE_s$ and $PRMSE_x$. See, for example, Haberman (2008b) or Sinharay and Haberman (2008) for further details on the PRMSEs and their computations.

Sinharay and Haberman (2008) and Sinharay (2010) found from a study of several operational and simulated data sets that augmented subscores lead to more accurate diagnostic information than subscores. For example, they showed that for several operational tests, the subscores do not have added value, but the augmented subscores do. Thus, for testing programs interested in reporting diagnostic scores based on classical test theory, the augmented subscores hold great promise.

However, while Puhan and Liang (2011) performed research on equating of subscores, there is a lack of research on equating of augmented subscores for the nonequivalent groups with anchor test design.

Methods

This section discusses our suggested methods for equating augmented subscores under the nonequivalent groups with anchor test design. Each of these methods is essentially the chain equipercentile equating (e.g., Kolen & Brennan, 2004) using suitably chosen anchor scores.[3] Here, the anchor score is a score that is on the same scale on both the new form sample and the old form sample and allows one to adjust for the difference in difficulty of the two test forms. Each of the equating methods proceeds as follows:

Table 1
The Anchor Scores Used by the Methods for Equating Augmented Subscores

Method	Anchor Score Used by the Method
1	Anchor subscore
2	Anchor total score
3	Anchor augmented subscore

- Given an augmented subscore on the new form, find its percentile rank in the new form sample.
- Find the anchor score that has the same percentile rank as above in the same sample.
- Find the percentile rank of that anchor score in the old form sample.
- Find the score on the old form that has the same percentile rank as in the above step in the old form sample.

The methods differ only in the way they define the anchor scores.

In the first method, to equate an augmented subscore one uses as an anchor score the score on the items that belong to the anchor test and also to the corresponding subarea—this score is referred to as the "anchor subscore." In the second method, to equate an augmented subscore one uses as an anchor score the total score on the anchor test. In the third method, to equate an augmented subscore one computes the corresponding augmented subscores from the anchor subscores—let us refer to these scores as *anchor augmented subscores*—and uses them as anchor scores. The anchor augmented subscore used in this method involves weights (or coefficients) placed on the anchor total score and the anchor subscore. One can fix the weights based on experience with previous administrations of the test or estimate the weights from the current sample. We employed the average of the weights from the new form sample and old form sample. For example, consider that the interest is in equating of the augmented subscore corresponding to subtest 1. We compute the weights on subscore 1 and the total score in augmented subscore 1 from the new form sample and then from the old form sample and then average them. These average weights are then imposed on the anchor subscores to compute an anchor augmented subscore. It is possible to explore other weights in future research. Table 1 shows the anchor scores used in the methods.

We also tried another method in which the total score (that is the sum of all subscores) on the new form is first equated to the total score on the old form using the anchor test. Then, to equate the augmented subscores, the investigator uses as anchor score the total score on the old form for the old form sample and the total equated score on the new form for the new form sample. Such an anchor was considered in Puhan and Liang (2011). This method performed exactly like the second method described above, so we do not discuss any results on this method henceforth.

Note that Methods 2 and 3 can be employed to equate the augmented subscore for a subarea that is not represented in the anchor test.[4] In addition, Methods 2 and 3 use, as anchor scores, not only the subscore of interest, but also the other subscores.

For example, in the equating of an augmented reading subscore, scores on listening items may also contribute to the anchor score. This may seem counter-intuitive. However, firstly, subscores are most often highly correlated with each other for operational tests (see, for example, Sinharay, 2010) and hence the anchor scores used in Methods 2 and 3 are highly correlated with the augmented subscores that are to be equated. Because the usefulness of an anchor increases as the correlation between the anchor score and the score to be equated increases (see, for example, Angoff, 1971; Petersen, Kolen, & Hoover, 1989), we can expect the anchor scores in these methods to lead to accurate equating. Secondly, Methods 2 and 3 have some similarity to the equating methods used in some testing programs such as the AP® examinations that involve the linking of a composite score, which is a weighted average of a multiple choice score and a constructed response score, to the multiple choice score (even though it is believed that constructed response items and multiple choice items often measure slightly different skills). Thirdly, for several tests involving constructed response items, equating using scores on multiple choice items as anchor scores is used or has been suggested (e.g., Livingston, 1994; Ercikan et al., 1998; Educational Testing Service, 2007). Livingston (1994), in a study using data from several operational tests, found that for tests with only a few constructed response items, the use of a related test with multiple choice items as an anchor for equating is clearly preferable to no equating.

Note that the weights (or coefficients) on the augmented subscores that are equated in this paper are determined by the above-mentioned formula from Haberman (2008b) and, as a consequence, the weights differ between the old form and the new form (that is, we may be equating ".38×Reading subscore + .12×Total score" on the new form to ".34×Reading subscore + .14×Total score" on the old form). It is possible to

- compute augmented subscores on several forms of a test, and then, if the weights are very close (or augmented subscores computed using different weights are highly correlated for each form), to
- fix the weights at a specific set of values (maybe at the average of the weights), and
- to equate the augmented subscores with fixed weights between forms (that is, equate ".35×Reading subscore + .13×Total score" on the new form to ".35×Reading subscore + .13×Total score" on the old form).

However, we do not consider equating of augmented subscores with fixed weights henceforth—this can be a potential area of further research. Augmented subscores with varying weights are the regression estimates of the corresponding true subscores and equating them makes more sense (so that essentially we are equating an estimate of the true subscore on the new form to an estimate of the true subscore on the old form). Augmented subscores with fixed weights, while convenient, will not have this property.

The above-mentioned methods can be used to equate any observed score that is a weighted linear combination of component scores so that the component scores measure different constructs.

Next, we apply the above-mentioned methods to a set of operational data. The goal will be to examine the accuracy of the methods and to compare the methods.

Application

Data

The original data for this example are from one form of a licensing test for prospective teachers. The test form included 119 multiple-choice items, about equally divided among four content areas: language arts, mathematics, social studies, and science. Scores on each of these content areas are reported—these are treated as the subscores here. The original form had been used at two test administrations and the two examinee samples play the role of the new form sample P and the old form sample Q in our analysis.

The item responses from the original test were used to construct two pseudo-tests, X (new form) and Y (old form). A pseudo-test consists of a subset of the test items from the original 119-item test, and the score on the pseudo-test for an examinee is found from the responses of that examinee to the items in the pseudo-test. The pseudo-tests X and Y each contain 44 items, 11 items from each of the four content areas: language arts, mathematics, social studies, and science. Tests X and Y have no items in common and were made 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 test. This anchor has no items in common with either X or Y. The mean percent correct on the anchor test approximately equaled that for the 119-item original test. The reliability in the combined sample $P + Q$ of the scores on the original 119-item test, X, Y, and the anchor are .91, .80, .80, and .75, respectively. Further details on the construction of these pseudo-tests can be found in von Davier et al. (2006). Because all the examinees in P and Q were administered all the 119 items on the original test, all of the examinees in P and Q have scores on X, Y, and the anchor. So, it is possible to compute the single-group equating function of the augmented subscores on X to the corresponding quantities on Y using the combined sample $P + Q$—this equating function will be treated as the *criterion equating function* for the augmented subscores.

The test X was constructed to be considerably easier than Y. For example, on Q, the mean score for X is larger than the mean score for Y by 133% of the SD of Y. In addition, Q is more able than P with a mean anchor-score that is higher than P by approximately a quarter of an SD in $P + Q$. The large difference in difficulty between X and Y was supposed to ensure that the equating functions would be nonlinear and the relatively large difference in the test performance of P and Q was supposed to ensure that different equating methods would produce different results for the pseudo-data.

Table 2 shows the average proportion correct scores of the samples P and Q for the four subscores on the original 119-item test. The table shows that the difference between the two samples is mostly similar, all lying between .016 and .043, across the four subscores. This is true even though the subscores belong to a variety of content areas. If the computations were performed on the anchor instead of on the original test, the pattern of differences would remain similar to that in Table 2.

Table 2
The Average Proportion Correct Subscores for P and Q for Application 1

Sample	Language Arts	Mathematics	Social Studies	Science
P	.685	.707	.657	.695
Q	.715	.750	.673	.734

Table 3
PRMSEs of the Subscores for Application 1

Subscore	$PRMSE_s$/Reliability	$PRMSE_x$	$PRMSE_{sx}$
1	.57	.75	.77
2	.62	.68	.73
3	.49	.67	.69
4	.62	.75	.78

The correlation coefficients among the subscores on X in P range between 0.38 and 0.53. Table 3 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 on X, or in Q instead of P, or both.

Table 3 shows that while the subscores have low reliability and hence do not have added value (because $PRMSE_s$ is less than $PRMSE_x$ for all subscores), the weighted averages have added value ($PRMSE_{sx}$ is somewhat larger than both $PRMSE_s$ and $PRMSE_x$ for all the subscores). Thus, this data set seems appropriate for applying the above-mentioned methods for equating of augmented subscores.

Equating of Augmented Subscores When The Anchor Test is Representative

To mimic the structure of the nonequivalent groups with anchor test design, we ignored the scores on X in Q and on Y in P, and performed the equating of augmented subscores using the methods described in the previous section using scores on X in P, Y in Q, and the anchor in P and Q.

The augmented subscores have fractional values. For example, the possible number of augmented subscores corresponding to content area 1 on X in P was found to be 408 for the data set; they are 2.43, 2.59, . . . 10.79, 10.95. It is possible to compute an equating function from X to Y for any real value between 0 (minimum possible score on the subtest) and 11 (maximum possible score on the subtest) of an augmented subscore. That is, unlike in usual equating scenarios, equating here is not restricted only to integer values. To simplify things, we compute the function equating augmented subscores only at integer values between 0 and 11 (note that the augmented subscores for X in P or Y in Q were not rounded before the equating). The anchor test was treated as an external anchor test although the methods discussed in this paper apply to internal anchor tests as well. Polynomial loglinear models (Holland & Thayer, 2000) were used to presmooth the marginal distributions of augmented

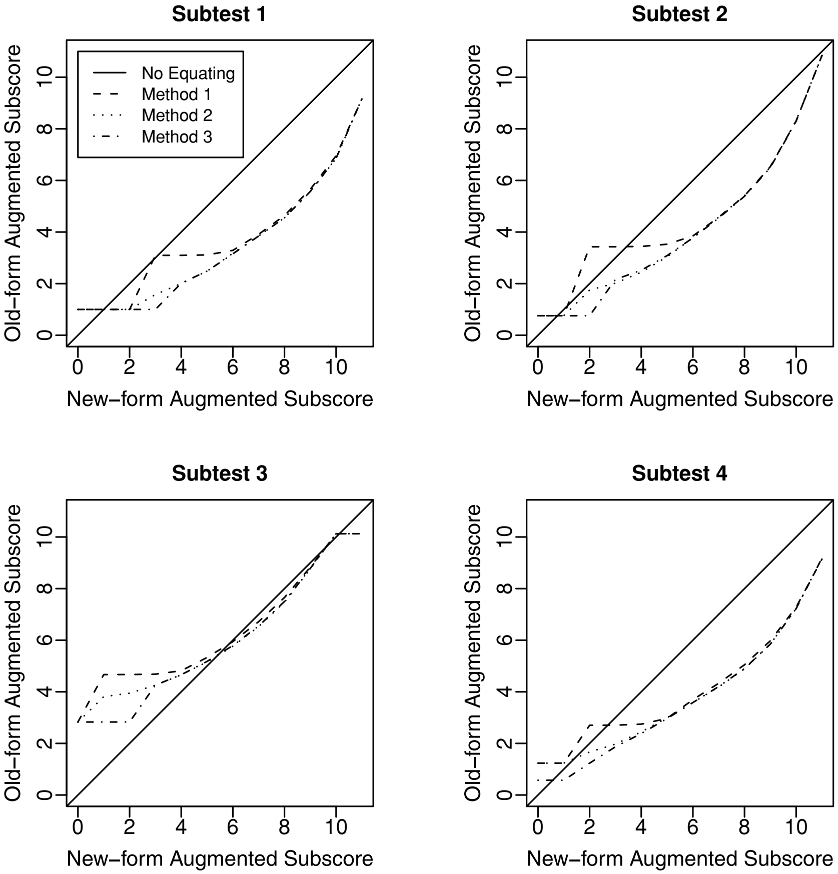

Figure 1. The equating functions for the augmented subscores for the application.

subscores and the anchor scores. While applying the above-mentioned methods, the linear interpolation method (Kolen & Brennan, 2004) was used to compute percentile rank functions corresponding to the discrete distributions of the total test scores, anchor subscores and anchor total scores. A version of the linear interpolation method that is appropriate for fractional scores was used to compute percentile rank functions corresponding to the discrete distributions of augmented subscores and the anchor augmented scores. For example, in applications of Method 1, the marginal distributions of the anchor subscores and the augmented subscores were presmoothed, the linear interpolation method was used to compute percentile rank functions corresponding to the discrete distribution of the anchor subscores, and a version of the linear interpolation method that is appropriate for fractional scores was used to compute percentile rank functions corresponding to the discrete distributions of the augmented subscores.

Figure 1 shows, for each content area, the equating function for the augmented subscores for all three methods. The horizontal axis shows the augmented subscore on form *X* and the vertical axis shows the corresponding equated augmented subscore

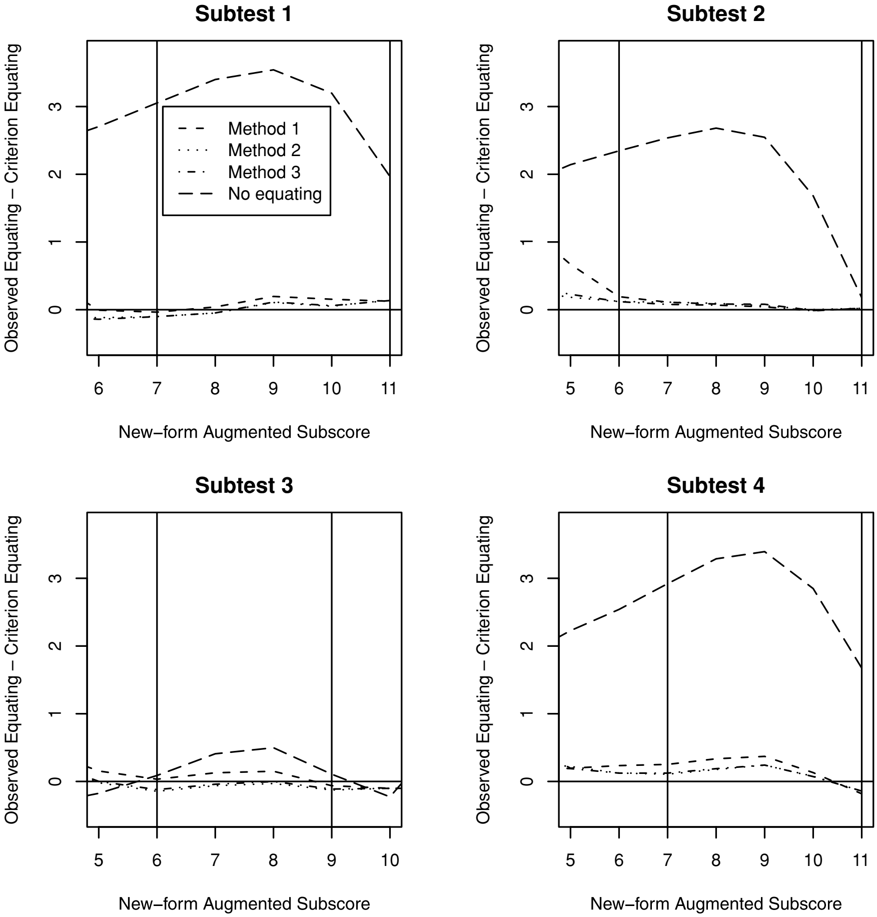

Figure 2. The differences between the observed and criterion equating functions for the application.

on form Y. The 45-degree line represents the identity equating, or, no equating.[5] There is a substantial difference in results between no equating and equating using Methods 1, 2, or 3 for subtests 1, 2, and 4. The three equating methods provide similar results except for the low augmented scores. However, the 1st percentiles for the four augmented subscores are 6, 5, 5, and 5—so there are very few examinees in the region where the equating methods substantially differ.

Figure 2 shows, for each of the three methods, the differences of the equatings of the augmented subscores and the criterion equating function. There is a solid horizontal line at 0; if the equating methods are accurate, the differences should be close to this line. The 5th and 95th percentiles of the augmented subscores in the combined sample $P + Q$ are also shown in the plot using solid vertical lines. The range of the X-axis in each panel is the 1st and 99th percentile of the corresponding augmented subscore in P. As a baseline for comparison, differences for no equating

Table 4
The RMSDs for the Methods for Equating Augmented Subscores for Application 1

Content	Representative Anchor			Nonrepresentative Anchor	No
Area	Method 1	Method 2	Method 3	Method 2	Equating
1	.15	.08	.09	.08	3.22
2	.14	.07	.06	.07	2.17
3	.14	.09	.08	.11	.38
4	.26	.16	.16	.18	2.93

Table 5
Correlation of the Augmented Subscore and Anchor Score in Q

Content Area	Method 1	Method 2	Method 3
1	.55	.74	.73
2	.69	.74	.76
3	.45	.74	.71
4	.52	.75	.74

is also added. The standard deviation of the augmented subscores on *X* range between 1.8 and 2.6 for the four subtests.

For each augmented subscore and method, we computed an overall measure of difference between an observed equating function and the criterion equating function by computing $\sqrt{\sum_i w_i(o_i - t_i)^2}$, where o_i is the observed equating function, t_i is the criterion equating function, and w_i is the weight at (sub)score point i (where $\sum_i w_i = 1$). Let us refer to this measure as the root mean squared difference (RMSD).

Columns 2–4 of Table 4 show the RMSDs for the four augmented subscores for the representative anchor. Column 6 of the table shows the RMSDs for no equating. To compute the weights w_is, the values of the augmented subscores on *X* in *P*+*Q* were rounded to the nearest integers and the weights were computed as proportional to the frequency of a rounded augmented subscore. The RMSD computed above is in units of augmented subscores (which are roughly of the same magnitude as the subscores). In equating studies, it is useful to have some idea about when an equating error is too large. Some authors, such as Dorans and Feigenbaum (1994), have elected to use .5 in the raw score scale as a rough guideline for an acceptable level of equating error. We will use the same guideline in this paper.

Figure 2 and Table 4 show that Methods 2 and 3 for equating augmented subscores perform very similarly and better than Method 1. Any method for equating augmented subscores is much better than no equating. The RMSD for no equating can be up to 3.22 whereas that for any of Methods 1–3 is at most .26.

Table 5 shows the correlation between the augmented subscores and the anchor score for sample *Q* for the three methods to equate augmented subscores.

Methods 2 and 3 are associated with higher correlations than Method 1 in Table 5. Thus, the better performance of Method 2 and Method 3 over Method 1 is expected given the conjecture that the higher the correlation between the anchor score and the score to be equated, the more useful is the anchor (see, for example, Angoff, 1971; Petersen et al., 1989).

Equating of Augmented Subscores When the Anchor Test is Nonrepresentative

In this analysis, we used the same pseudo-tests X and Y as above, but used, instead of the 24-item anchor test mentioned above, a shorter 12-item anchor test that has the same six items each as in the 24-item anchor from the first two content areas and no items belonging to the last two content areas. In other words, the anchor test can be considered nonrepresentative of the tests to be equated in this situation. We used Method 2 for equating augmented subscores. Note that Method 1 cannot be used for the last two content areas—so we did not use this method here. Column 5 of Table 4 shows the RMSDs for the equating of augmented subscores for Method 2 with a nonrepresentative anchor. The RMSD values show that the equating method leads to accurate results and performs much better than no equating even for the nonrepresentative anchor. The RMSDs are only slightly worse than those for the representative anchor. Also, the RMSDs for the nonrepresentative anchor for the last two content areas are small even though the anchor test does not contain any items belonging to these two content areas. This is possible because the anchor total scores perform well as anchor scores, that is, the difference in the average anchor total score between P and Q reflects the difference of each average subscore between P and Q (which is clear from a look at Table 2, which shows that the difference between P and Q is similar across the four subareas).

The length of the anchor test (of 12 items) in this case is less than that recommended by most equating experts (see, e.g., Kolen and Brennan, 2004), but equating with such a short anchor test is better than no equating.

Sinharay and Haberman (2011) analyzed data from pseudo-tests created from another form of the same test. The difference between the two tests X and Y and the difference between two samples P and Q were smaller for these data than the application described above. The results were similar. For example, Methods 2 and 3 perform better than Method 1 and the RMSDs of equating of augmented subscores with nonrepresentative anchors are small.

Simulation Study

Although the two operational data examples provided us with some idea about the performance of our suggested methods, they represent only a tiny fraction of all possible equating scenarios. For example, they do not involve subtest length of more than 12. Therefore, we performed a simulation study to examine the performance of the above-mentioned equating methods under several scenarios. The design of the study is somewhat similar to those in Sinharay and Holland (2007) and Sinharay (2010).

Simulation Design

We obtained a data set from the licensing test referred to in the application. The 120 multiple choice items on the test contribute to the above-mentioned four subscores. Because the four subscores can be considered to measure four different but correlated dimensions, we fitted to the data set a multidimensional IRT model (MIRT; e.g., Reckase, 2007) with 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} denote slope parameters and b_i denote 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 among a_{1i}, a_{2i}, a_{3i}, and 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 while $a_{2i}=a_{3i}=a_{4i}=0$), so that the MIRT model has a simple structure.

The estimated item parameter values from the fitting of the model given in (1) to the data set were instrumental in obtaining the generating item parameters of tests *X* (new form), *Y* (old form) and the anchor in the simulations. A bivariate normal distribution was fitted to the log-slope and difficulty parameter estimates corresponding to each content area (or subscore). The generating item parameters for a content area of *Y* were randomly drawn from the corresponding bivariate normal distribution. To compute the generating item parameters for a content area of *X*, we made random draws from the corresponding bivariate normal distribution and then added .25 to the difficulty parameter component of all the draws. This strategy ensured that the test *X* was slightly more difficult than test *Y* and the difference was similar over all the content areas.

We assume that the anchor test length is about 40% of that of *X* or *Y*. To compute the generating item parameters for a content area for the anchor test, we made random draws from the corresponding bivariate normal distribution, and then added a constant .125 to the difficulty parameter component of all the draws. This strategy ensured that the difficulty of the anchor is the average of the difficulties of forms *X* and *Y*. The generating item parameters for the tests *X*, *Y*, and the anchor were the same for all 100 replications for a simulation condition.

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

(i) Length of the subtests. This paper used three values for the length: 12, 20, and 30. Note that the reliability of a test increases as the test length increases. For simplicity, this paper assumes that the different subtests of a given test have the same length.

(ii) Level of correlation (ρ) among the components of $\boldsymbol{\theta}$. This paper used three levels: .70, .80, and .90. A survey of operational data in Sinharay (2010) shows that this is a realistic range. If the correlation level for a simulation case is ρ, the mean of all the off-diagonal elements of Σ (which denote 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

$$\mathcal{C} = \begin{pmatrix} 1.00 & .78 & .80 & .84 \\ 0.78 & 1.00 & .72 & .78 \\ 0.80 & .72 & 1.00 & .88 \\ 0.84 & .78 & .88 & 1.00 \end{pmatrix},$$

which is the estimated correlation matrix between the components of $\boldsymbol{\theta}$ from the fit of the model given by Equation 1 to the above-mentioned licensing test data set. To obtain a Σ with a mean correlation ρ, the mean of the correlations of $\mathcal{C}$, denoted henceforth as m, was computed and then the (i, j)th element of Σ was set as $\rho - m$ + the (i, j)-th element of $\mathcal{C}$, where $i \neq j$.[6] This strategy ensured that the average of the correlations in Σ is ρ, but allowed the correlations between the subscores to be realistically different.

The steps of the simulation. For each simulation condition (determined by a "correlation" and a "subtest length"), the generating item parameters of the X-test, the Y-test, and the anchor test were randomly drawn (as described earlier) once, and then 100 replications were performed. The sample size of both P and Q was set to 2,000. Each replication involved the following three steps: (1) generate the ability parameters, (2) simulate the scores, and (3) perform equating. The details of the steps are provided below:

(i) Generate the ability parameters $\boldsymbol{\theta}$ for the samples 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 above to ensure that the average correlation is ρ. We used $g_P(\boldsymbol{\theta}) = \mathcal{N}_4(\boldsymbol{\mu}_P, \Sigma)$, where $\boldsymbol{\mu}_P$, which quantifies the difference between P and Q, is given by $\boldsymbol{\mu}_P = (0.25, 0.25, 0.25, 0.25)'$, i.e., the difference between the old form and new form samples is the same for all components of the ability and the difference is a borderline extreme ability difference that is rarely surpassed for large-scale operational tests. Table 2 shows that this choice of $\boldsymbol{\mu}_P$ is reasonable for a test like the one considered in the application.

(ii) Simulate scores on X in P, Y in Q, and those on the anchor in both P and Q from the MIRT model given by Equation 1 using the draws of $\boldsymbol{\theta}$ from step 1 and the generated item parameters for X, Y, and the anchor.

   (a) Perform equating of augmented subscores using the three methods using the scores of X in P, Y in Q, and the anchor in P and Q. To simplify matters, equating of augmented subscores was computed only for the possible values of the corresponding subscore (that is, for a simulation case with 12-item subtests, the equating was performed for 0, 1, ...12).

   (b) Ignore the scores on the items on the anchor test that belong to the third and fourth content areas and equate the augmented subscores using Method 2 using the items on the anchor test that belong to content areas 1 and 2. This represents the nonrepresentative anchor test case that was considered in the application.

Computation of the performance criteria: equating bias, SD, and RMSE. After the equating results from the 100 replications were obtained, to judge the performance of the equating, we computed bias (a measure of systematic error in equating) and SD (a measure of random error in equating) as performance criteria. For a simulation condition, let $\hat{e}_i(x_s)$ be the equating function in the ith replication for an augmented subscore x_s from X to Y. Suppose $e(x_s)$ denotes the value of the corresponding true/population equating function. The appendix section describes the computation of the true equating function. The bias at (sub)score-point x_s is then obtained as

$$\text{Bias}(x_s) = \frac{1}{100}\sum_{i=1}^{100}[\hat{e}_i(x_s) - e(x_s)] = \bar{\hat{e}}(x_s) - e(x_s), \text{ where } \bar{\hat{e}}(x_s) = \frac{1}{100}\sum_{i=1}^{100}\hat{e}_i(x_s).$$

The corresponding standard deviation is obtained as

$$\text{SD}(x_s) = \left\{\frac{1}{100}\sum_{i=1}^{100}[\hat{e}_i(x_s) - \bar{\hat{e}}(x_s)]^2\right\}^{\frac{1}{2}}.$$

The corresponding root mean squared error (RMSE) is computed as

$$\text{RMSE}(x_s) = \left\{\frac{1}{100}\sum_{i=1}^{100}[\hat{e}_i(x_s) - e(x_s)]^2\right\}^{\frac{1}{2}}.$$

It can be shown that

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

i.e., the RMSE combines information from the random and systematic error.

The overall performance of a method for a simulation case can be judged by the overall (or "weighted sum of") bias, $\sum_{x_s} r(x_s)\text{Bias}(x_s)$, the overall SD, $\sqrt{\sum_{x_s} r(x_s)\text{SD}^2(x_s)}$, and the overall RMSE, $\sqrt{\sum_{x_s} r(x_s)\text{RMSE}^2(x_s)}$.

Simulation Results

Table 6 shows the following summary statistics for the several simulation conditions:

- Average reliability of the total score (α)
- Average correlation between the subscores ($\bar{r}$)
- Average of the PRMSEs ($PRMSE_s$, $PRMSE_x$, and $PRMSE_{sx}$)
- Overall percent of subscores that have added value (% sub)
- Overall percent of augmented subscores that have added value (% wtd)

These summary statistics were computed using the X-test from all 100 replications. For example, the value .61 of $PRMSE_s$ for subtest length 12 and correlation

Table 6
Summary Statistics from the Simulation Study

		Correlation		
Subtest Length	Quantity	.70	.80	.90
12	α	.82	.84	.85
	$\bar{r}$	.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
	$\bar{r}$	.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
	$\bar{r}$	.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

.70, is the average of the 400 subscore reliability values (because there are four subscore reliability values in each of the 100 replications) for that simulation case. For another example, the value of 25 for "% sub" for subtest length 12 and correlation .70 indicates that out of the 400 simulated subscores (four in each of the 100 replications) for the simulation case, 25%, that is, 100 were of added value.

Figure 3 shows the average RMSEs for all simulation cases with $\mu_P = (0.25, 0.25, 0.25, 0.25)$. The simulation case is shown along the X-axis. The first three columns of symbols in Figure 3 show the RMSEs for subtest length 12, the next three for subtest length 20, and the last three for subtest length 30. For any subtest length, the three columns show the RMSEs for $\rho = 0.7$, .8, and .9, respectively. For any subtest length and ρ, Figure 3 shows for each method the average of the RMSEs of the four subscores.[7] For convenience, the order in which the methods appear in the box near the top left corner of the figure is roughly that of their average RMSEs (for example, "No equating," which has the largest average RMSE, appears first in the box). Sinharay and Haberman (2011) provide more details on the results. In interpreting the RMSEs in the figure, note that as test length increases (from subtest length 12 to subtest length 30), the standard error of measurement also increases (see, for example, Lord, 1959) and hence the RMSEs are also expected to increase. The

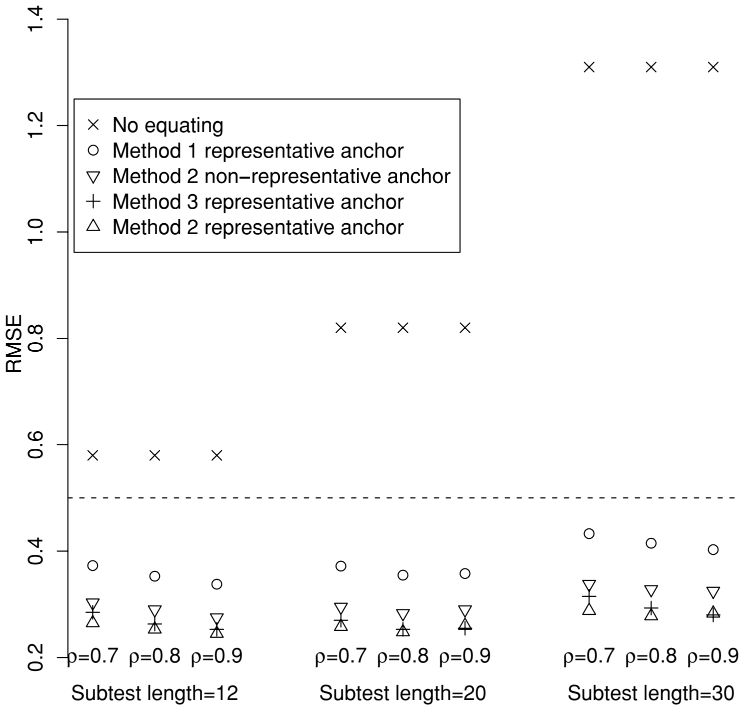

Figure 3. The Average RMSEs for $\boldsymbol{\mu}_P = (0.25, 0.25, 0.25, 0.25)$.

difference in RMSE between the methods arises mostly from a difference of bias; the SDs of the three methods are usually very close to each other for all simulation cases. Sinharay and Haberman (2011) provide the values of bias and SD for these simulation cases. The rough guideline for an acceptable level of equating error (Dorans & Feigenbaum, 1994), which is .5 in this case, is shown using a dashed horizontal line.

Table 6 and Figure 3 show the following:

- In Table 6, the percent of the augmented subscores that have added value is most often larger than 50 while the percent of subscores that have added value is most often less than 50. Hence, these simulation cases are appropriate for discussing equating of augmented subscores.
- The equating methods perform accurately. The RMSE of all the methods are small, that is, they are less than the rough guideline for an acceptable level of equating error. This is in agreement with the results observed earlier for the operational data examples.
- Any method leads to a substantial improvement over no equating for all simulation cases. Figure 3 shows that the RMSE for no equating is more than the rough guideline for an acceptable level of equating error for all simulation cases, whereas the RMSE for any equating method, even that using nonrepresentative anchor, is always less than the rough guideline for an acceptable level of equating error. So the use of any of the suggested methods will most often lead to a

more precise reported score rather than no equating.

- For equating with a representative anchor, both Methods 2 and 3 perform better than Method 1. Hence, borrowing information from other content areas to form an anchor score helps in equating of the augmented subscores. Among Methods 2 and 3, the former performs slightly better than the latter overall—the RMSE is more often smaller for the former than the latter.
- As the correlation level increases, the RMSE for any method tends to become slightly smaller.
- RMSE is larger for the longer subtests than for shorter subtests.
- The quality of equating with nonrepresentative anchors is mostly worse than that with representative anchors. However, equating is accurate even for the simulation cases when the anchor test is not representative. The small error of equating with nonrepresentative anchors in the simulations is in agreement with the results from the operational data sets. One reason behind this is that the anchor test should ideally reflect the differences in the two samples on the scores to be equated; and because of the assumption of equal difference for all content areas in the simulation above, the anchor total score, which is used as the anchor score in Method 2, reflects the differences in the two samples accurately.

Additional Simulations

The simulations above were performed under the assumption that $\boldsymbol{\mu}_P = (0.25, 0.25, 0.25, 0.25)'$, that is, the difference in the ability of the two samples is the same for all the subscores, and that the difference in the difficulty of the two tests is the same for all the content areas. Sinharay and Holland (2007) discussed other possible patterns of differences between the two samples and the two tests when data are generated from a MIRT model, and it is important to examine whether the results reported in Figure 3 hold under those patterns.

Hence, additional simulations were performed under other conditions. Varying patterns of difference in difficulty of the two tests did not affect results—so the pattern was set as in the earlier simulations, that is, X is more difficult than Y in all the content areas. However, varying patterns of differences in ability of the two samples substantially affected the results—so we report results for two of these patterns.

Figure 4 shows average RMSEs for the case when $\boldsymbol{\mu}_P = (0.1, 0.15, 0.2, 0.25)$, which is likely to occur in practice, as can be seen from Table 2. Results for $\rho = 0.7$ and .9 are shown in the figure.

Figure 5 shows the average RMSEs for the cases when $\boldsymbol{\mu}_P = (0.1, 0.1, -0.1, -0.1)$. This kind of pattern, where P is better than Q in some content areas but worse in some others, was discussed in Klein and Jarjoura (1985). See, for example, their Figures 2 and 3.

For equating with a nonrepresentative anchor under this choice of $\boldsymbol{\mu}_P$, the RMSEs for the first two subscores were considerably different from RMSEs for the last two subscores—so Figure 5 shows the average RMSEs for Subscores 1–2 and Subscores 3–4 separately.

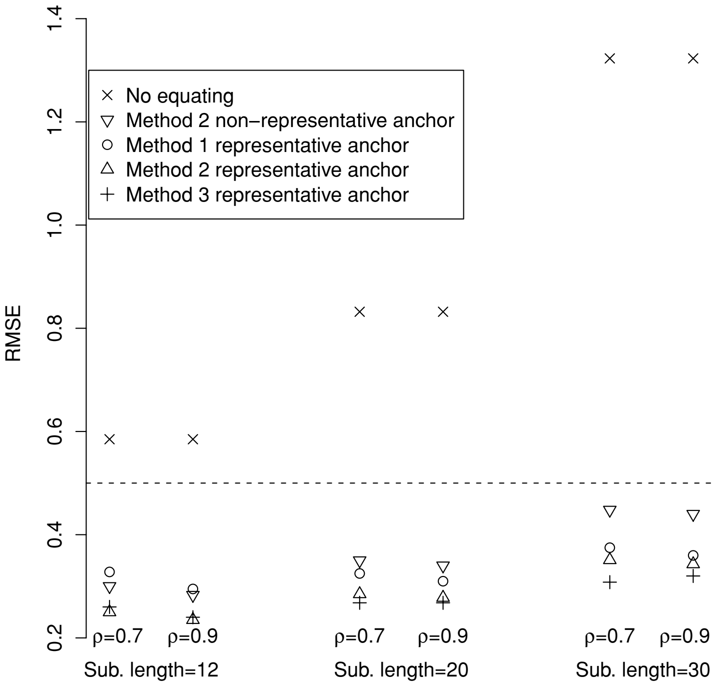

Figure 4. The Average RMSEs for $\boldsymbol{\mu}_P = (0.1, 0.15, 0.2, 0.25)$.

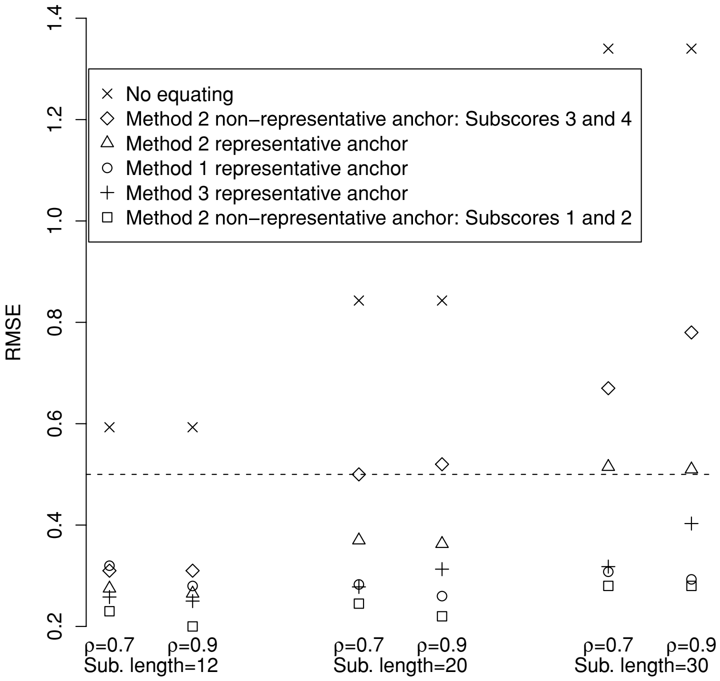

Figure 5. The Average RMSEs for $\boldsymbol{\mu}_P = (0.1, 0.1, -0.1, -0.1)$.

Figures 4 and 5, the results of which are often similar, draw a slightly different picture regarding the performance of the suggested equating methods from that drawn by Figure 3 because Figures 4 and 5 show that:

- The average RMSEs are often much larger in Figure 5 in comparison to those in Figure 3. For example, the largest average RMSE in Figure 3 for any equating method (except, of course, "No equating") was about .45, whereas the RMSE can be as large as about .80 in Figure 5. Even then, equating with some method is better than no equating according to Figures 4 and 5. Also, the average RMSE is most often less than the rough guideline for an acceptable level of equating error (0.5) in these figures.
- Unlike in Figure 3, Method 2 for equating augmented subscores is often the worst of the three methods for the representative anchor case, especially in Figure 5. In the simulations that resulted in Figures 4 and 5, the difference on the anchor total score between the two samples does not reflect the difference on any single subscore between the two samples. For example, when $\mu_P = (0.1, 0.1, -0.1, -0.1)$, which led to Figure 5, the average anchor total scores for the two samples are close (because each sample is stronger than the other in two subareas and weaker in two subareas) while the average of any subscore differs by one-tenth of a standard deviation between the two samples. Hence, for the simulation cases represented in Figures 4 and 5, the anchor total score does not perform well as an anchor.
- Unlike in Figure 3, a nonrepresentative anchor often leads to much worse equating of subscores compared to a representative anchor in Figures 4 and 5. For example, the average RMSE for the last two subscores for correlation = 0.90 and subtest length = 30 for $\mu_P = (0.1, 0.1, -0.1, -0.1)$ is about .8 for the nonrepresentative anchor case compared to values of .3–0.5 (for Methods 1–3) for the representative anchor case. This is due to the fact that the anchor total score, which acts as the anchor score in Method 2 of the nonrepresentative anchor case, is based on only the first two subareas. Because P is stronger than Q in these two subareas (as the first two components of μ_P are positive), P appears stronger than Q according to the anchor score in this case; however, P is weaker than Q on the third and fourth subareas. So the anchor score does not properly reflect the difference between P and Q and substantial equating error is expected in equating of the third and fourth augmented subscores. The equating error for the first two augmented subscores is smaller for the nonrepresentative anchor case than the representative anchor case because, unlike in the latter case, the anchor total score in the former case properly reflects the difference between the performances of the two samples on these two subareas.

The results above were obtained when the anchor test length was 40% of the total test length. We computed average RMSEs for the case when the total test length is 120 (30 items per subarea), the anchor test length is 24 (that is, 6 items per subarea), and $\mu_P = (0.25, 0.25, 0.25, 0.25)$. Figure 6 shows these average RMSEs along with the average RMSEs for the 30-item-per-subtest cases of Figure 3.

Figure 6 shows that while the average RMSEs for all methods increase when the total-test-length-to-anchor-test-length decreases from 40% to 20%, the average

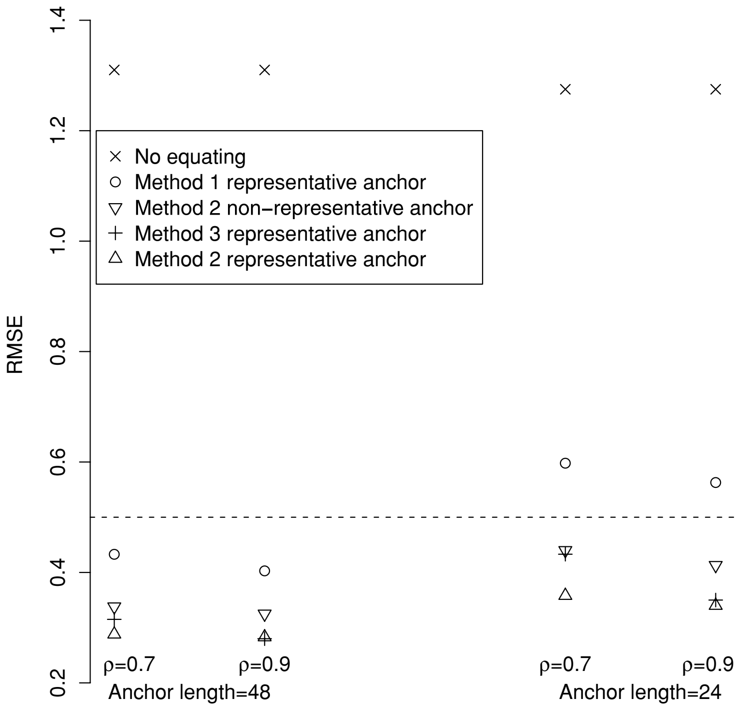

Figure 6. The RMSEs for the 24-item-anchor and 48-item-anchor when the number of items per subtest is 30 and $\boldsymbol{\mu}_P = (0.25, 0.25, 0.25, 0.25)$.

RMSE for Method 1 increases most. That is because Method 1 does not allow borrowing of information from other content areas to form the anchor score, which is crucial for short anchor tests.

Conclusions

We suggested several methods for equating of augmented subscores under the nonequivalent groups with anchor test design and examined the performance of the methods using several operational and simulated data sets. The results demonstrate that the suggested methods perform quite accurately, that is, the extent of equating error (both systematic and random) is small under most practical situations. Even when the anchor test is nonrepresentative of the test (that is, some content areas covered in the test are not covered in the anchor test), the suggested methods often perform quite well. The results also demonstrate that borrowing information from other subscores to compute an anchor score often leads to more accurate equating of augmented subscores—that is why Methods 2 and 3 mostly perform better than Method 1. The simulations show that the equating error may be large if the difference between the two samples is dissimilar over the different content areas. Methods 2 and 3 seem to perform worse than Method 1 in such cases, that is, borrowing information from other subareas to form the anchor score is harmful. A big question then is: "How often are the differences between examinee samples dissimilar over different content areas?" We examined, for one form each from four licensure tests[8] (three of which were considered in Puhan et al., 2010), the proportional correct subscores for several examinee samples (where a sample represents all the examinees who took the test form on a specific date). It was found, as in Table 2, that the difference between examinee samples is mostly similar over the different subareas, which means that on most occasions the simulation results reported in Figure 3 will hold, all the methods for equating weighted averages will perform accurately, and Methods 2 and 3 will perform better than Method 1.

The usual limitations of simulation studies apply to the simulation results reported in this paper. However, our simulations are somewhat realistic because (i) the results for the operational data sets essentially agree with those from the simulation study, (ii) we used item parameters estimated from operational data to simulate the data sets, and (iii) MIRT models usually fit test data better than a univariate IRT model (e.g., Haberman, von Davier, & Lee, 2008)—so the data simulated using a MIRT model in this study are expected to resemble operational data reasonably well.

There are several issues that can be examined further. For example:

- More operational data sets and simulated data sets, especially those different in nature from those considered in this paper, should be analyzed using the methods suggested in this paper.
- Population invariance of equating (e.g., Dorans & Holland, 2000) of augmented subscores is a potential area of future research.
- This paper considered chain equipercentile equating. Other equating methods, for example, kernel equating (e.g., von Davier, Holland, & Thayer, 2003) and equating using continuous exponential families (Haberman, 2008a) could also be applied.

Appendix: Computation of the True/Population Equating Function

The true/population equating function for the augmented subscores is assumed to be the same as the true equating function for the subscores. That is because the augmented subscore is an estimate of the true subscore (see, e.g., Haberman, 2008b) and hence the equating function of the augmented subscores can be considered to be an estimate of the true equating function of the subscores.

The true equating function for any subscore for a simulation case can be seen 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 Test X and the corresponding subscore Y_s on Test 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. We used the iterative approach of Lord and Wingersky (1984) to obtain $P(X_s = x_s|\boldsymbol{\theta})$, the probability of obtaining a raw subscore of $X_s = x_s$ by an examinee with ability $\boldsymbol{\theta}$. This approach involves the values of the item parameters—we used the generating item parameters for the items contributing to X_s. Once $P(X_s = x_s|\boldsymbol{\theta})$ is computed, $r(x_s)$, the probability of a raw subscore of x_s on test X_s in population T is obtained by numerical integration as

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

where $g_T(\boldsymbol{\theta}) = 0.5g_P(\boldsymbol{\theta}) + .5g_Q(\boldsymbol{\theta})$. Because we assumed a simple structure, for any subscore s, $P(X_s = x_s|\boldsymbol{\theta})$ is a function of the corresponding component of $\boldsymbol{\theta}$, that is, θ_s, and the above integration ends up being a one-dimensional integration, 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 test 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 as $R(x_s)$ and $Q(y_s)$ respectively. The true equating function for subscore s is then obtained as

$$\text{eq}(x_s) = Q^{-1}(R(x_s)).$$

The true equating function is the same for each replication and correlation level, but varies with "subtest length."

Acknowledgments

The research was funded by Educational Testing Service. The authors thank Brian Clauser, Gautam Puhan, Alina von Davier, Danier Eignor, Samuel Livingston, and the anonymous reviewers for their helpful advice. The authors gratefully acknowledge the help of Kim Fryer with proofreading. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Notes

[1] Sinharay (2010) showed that the linear combination of Haberman (2008b) is almost always very close to the augmented subscore of Wainer et al. (2000) and any method for equating the linear combinations of Haberman (2008b) suggested in this paper will apply to the augmented subscores of Wainer et al. (2000)—so referring to the linear combinations of Haberman (2008b) as augmented subscores (Wainer et al., 2000) is not an abuse of the name.

[2] Rounding the augmented subscores is a way around that, but that may lead to inaccurate equating, especially for short to medium-length tests.

[3] Though we discuss only chain equipercentile equating in this paper, a linear equating method such as the chain linear method or Tucker method (e.g., Kolen & Brennan, 2004) can be applied instead, especially if the sample size is small.

[4] Of course the accuracy of such equating has to be checked before implementing it.

[5] This line is shown because several testing programs operationally report subscores that are not equated.

[6] When the level of correlation is .90, the (3,4)th and (4,3)th element of $\mathcal{C}$ were changed to .85 before this calculation to ensure that Σ is positive definite.

[7] The RMSEs rarely differed much over the four subscores—see Sinharay and Haberman (2011) for the RMSE for each subscore—so averaging them does not lead to much loss of information.

[8] The test considered in the Application section is not one of these.

References

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

Chen, H., Lu, T., Thayer, D. T., & Han, N. (2007). *LOGLIN/KE version 3.0*. Princeton, NJ: Educational Testing Service.

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

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.

Educational Testing Service (2007). *General science: Content essays (0433), test analysis form 4CPX1* (Unpublished statistical report No. SR-2007-109). Princeton, NJ: Author.

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.

Haberman, S. J. (2008a). *Continuous exponential families: An equating tool* (Research Report No. 08-05). Princeton, NJ: Educational Testing Service.

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

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

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, NY: Springer.

Livingston, S. A. (1994). *Equating constructed-response tests through a multiple-choice anchor: A small-scale empirical study* (Unpublished statistical report No. SR-94-100). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, *19*, 233–239.

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Puhan, G., & Liang, L. (2011). Equating subscores under the non-equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice, 30*(1), 23–35.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266–285.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Amsterdam, The Netherlands: 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* (Research Memorandum No. 08-18). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Haberman, S. J. (2011). *Equating of subscores and weighted averages under the NEAT design* (Research Report No. 11-01). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, 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* (Research Report No. 06-02). Princeton, NJ: Educational Testing Service.

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

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Authors

SANDIP SINHARAY is a Principal Research Scientist, Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541; ssinharay@ets.org. His primary research interests include item response theory, equating, diagnostic score reporting, Bayesian methods, and application of statistics to education.

SHELBY J. HABERMAN is a Distinguished Presidential Appointee, Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541; shaberman@ets.org. His principal research interests are analysis of qualitative data, asymptotic approximations, and application of statistics to educational measurement.