Effectiveness of Collateral Information for Improving Equating in Small Samples

Sooyeon Kim,
Samuel A. Livingston
Charles Lewis
Educational Testing Service
Princeton NJ

Paper presented at the annual meeting of the
American Educational Research Association (AERA) and the
National Council on Measurement in Education (NCME)

April 13-17, 2009, San Diego, CA.
Abstract

This study examined the effectiveness of using collateral information to improve the accuracy of equating in small samples of examinees. Collateral information from the equating of other tests was incorporated into an empirical Bayes (EB) estimate of the reference-form score corresponding to each new-form raw score. The evaluation consisted of resampling procedures using data from large-group equatings of 10 different test forms, representing 9 different tests. Each large-group equating provided the data for repeated small-sample equatings based on resampling. The small-sample equatings were done by 5 methods, including 2 EB methods which used the equating of the other 9 test forms as collateral information. The small-sample equating methods were evaluated for agreement with the large-group equating results. The new-form sample size in the small-sample equatings was systematically varied (10, 25, 50, 100, 200) as was the sample size in the equatings used as collateral information (100, 200, 1400+). The results indicated that the use of collateral information tended to improve the accuracy of the equating when the new-form sample size was 25 or fewer.
Introduction

Test score equating is intended to provide examinees with fair and accurate scores by establishing effective equivalence between scores on two test forms that are designed according to the same specifications but contain different items. As with other statistical procedures, the equating of test scores is subject to sampling effects. If the sample of examinees is large and representative of the population, the equating relationship in the sample is likely to represent accurately the equating relationship in the population. The smaller the sample, the more likely it is that the equating function computed in that particular sample will differ substantially from that in the population.

Test equating often must be performed with a small sample, e.g., 20-30 people taking one of the test forms. The accuracy of the equating depends on the size of the samples, as has been shown theoretically, in the formulas for the standard error of equating (Kolen & Brennan, 2004), and empirically, in resampling studies (Kim, von Davier, & Haberman, 2008; Parshall, Houghton, & Kromrey, 1995; Skaggs, 2005). Nevertheless, many practitioners must use small data sets to determine the equating transformation to be used in an operational testing situation, because many testing programs must report comparable scores for a new edition of an established test form in a timely manner, regardless of sample sizes.

Using Collateral Information

One way to improve the accuracy of an equating based on small samples might be to use collateral information from other equatinings. Livingston and Lewis (2009) proposed an empirical Bayes (EB) estimation procedure that is a compromise between a sample equating function and the average of several prior equatins. In their EB procedure, this procedure is
applied separately at each possible raw score on the targeted new form (i.e., the new form to be equated).

EB methods involve the surprising result that estimates of a given population parameter can be improved by including in the estimation procedure information about other populations. An example for professional baseball given in an article by Efron and Morris (1977) on Stein's paradox in statistics will help to illustrate the fundamental idea underscored in the EB method. James and Stein (1961) showed that when more than two averages must be simultaneously estimated, the arithmetic mean becomes inadmissible as an estimator in the sense that there are other estimators that are superior to the mean for all values of the population parameter. The essential process in the James-Stein method is the “shrinking” of all the individual averages toward the grand average. It has been known that the optimal estimate from a set of data is the mean. What Stein showed was the overall mean is a more stable estimate than the individual means are, so incorporating the overall mean to some extent will provide greater stability to the individual estimates. In some cases, the estimate will be less accurate due to substantial bias caused by the improper shrinking. On average, however, the gain in precision more than offsets the bias. It is paradoxical because deliberately introducing bias can sometimes improve the estimates.

In mathematical notation, the equated score – the reference-form raw score corresponding to a given raw score on the new form – is computed by

$$
\hat{Y}_{EB} = \frac{1}{\hat{\sigma}^2_{prior}} Y_{prior} + \frac{1}{\hat{\sigma}^2_{eq}} Y_{eq} = \frac{\hat{\sigma}^2_{eq} Y_{prior} + \hat{\sigma}^2_{prior} Y_{eq}}{\hat{\sigma}^2_{eq} + \hat{\sigma}^2_{prior}},
$$

(1)
where $y_{eq}$ indicates the value implied by the current (small-sample) equating, $y_{prior}$ indicates an estimate based on the collateral information, $\hat{\sigma}_{eq}^2$ represents the variance of the equated score (i.e., the square of the conditional standard error of equating), and $\hat{\sigma}_{prior}^2$ represents the variance of the estimate based on collateral information. That variance is estimated by

$$\hat{\sigma}_{prior}^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left( y_i - \bar{y} \right)^2 - \frac{1}{m} \sum_{i=1}^{m} \hat{\sigma}_i^2, \tag{2}$$

where $i$ indexes the equatings, with $i = 1$ for the current equating and $i = 2$ to $m$ for the prior equatings, $y_i$ is the value observed in equating $i$, $\hat{\sigma}_i^2$ is its estimated variance over repeated sampling of examinees, and $\bar{y}$ is the mean of the observed $y_i$ values. In case $\hat{\sigma}_{prior}^2$ is negative, this variance will be replaced by zero. Under such a situation, $\hat{y}_{EB}^\prime$ will be the same with $y_{prior}$. The current equating is included in the set of equatings used to estimate the prior mean and variance, because the prior mean and variance are estimates for a domain of equatings, and the current equating is a member of that domain. The inclusion of the current equating in the estimation of the prior mean and variance increases the influence of the current equating on the posterior equated score ($\hat{y}_{EB}^\prime$), particularly when very few prior equatings are used as collateral information.

Notice in Equation 1 that the weight for the equated score observed in the current equating ($y_{eq}$) is the inverse of its variance. That variance decreases as the size of the equating samples increases. Therefore, the current equating will receive a greater weight as the size of its equating samples increases – particularly the smaller sample, which is usually the new form sample. The weight for the equated score implied by the collateral information
$(y_{prior})$ is the inverse of its variance. Equation 2 shows that this variance is large when the individual equatings used as collateral information produce very different results, especially if the variance of those individual equatings is small.

This point-by-point EB procedure does not require the equating transformation to follow a particular mathematical form. It treats the determination of the equated score at each point as a separate estimation problem to be solved by incorporating information from prior equatings. Bayesian estimates are derived separately for each equated score, rather than for the parameters of the equating function, allowing the procedure to estimate nonlinear equating functions without assuming that they have a particular mathematical form. Extensions of the procedure permit the use of collateral information from the equating of test forms having different numbers of items, by expressing the raw score on each form as a percentage of the maximum possible raw score on that form (i.e., percent-correct). If the new form used to establish the collateral information has a different number of items than the targeted new form, the percent-correct scores possible on the targeted new form will be different from those possible on the prior form. In that case, interpolation is necessary to determine the equated scores in the prior equating. The general EB procedures are described elsewhere (Kim, Livingston, & Lewis, 2008).

An important question in the use of this procedure is whether to include collateral information from the equating of forms of other tests. Limiting the domain to the forms of a single test can narrow the field of prior equatings down to a small sample that may not be representative of a domain that includes equatings of all future forms of the test. The previous forms of a single test may have been much more alike in difficulty than the future forms will be. In this case, limiting the domain of prior equatings to forms of that test would lead to an
underestimate of \( \text{var}(y_{\text{prior}}) \). The resulting adjustment formula would give too much weight to \( y_{\text{prior}} \) and too little weight to \( y_{eq} \).

Recently, Kim, Livingston, and Lewis (2008) conducted two resampling studies to determine empirically how much the accuracy of small sample equating could be improved by using EB methods with collateral information from equatings of previous forms of the test to be equated. Bayesian and non-Bayesian equating procedures were applied to samples ranging in size from 10 to 200 observations drawn from large groups of examinees who had taken the tests in a non-equivalent groups with anchor test (NEAT) equating design. The procedure was applied separately to the equating of two different forms of the same test, incorporating nearly the same collateral information in both cases. New Form 1 was similar in difficulty to its reference form, as were most of the forms used as collateral information. The use of collateral information from the other forms improved the accuracy of the small-sample equating of New Form 1. New Form 2, unlike the forms in the equatings used as collateral information, was much more difficult than its reference form. Consequently, the use of collateral information made the equating of New Form 2 slightly less accurate. The increase in accuracy from using collateral information in the equating of New Form 1, however, was much greater than the decrease in accuracy from using collateral information in the equating of New Form 2. In both cases, the use of the collateral information created a bias in the equating of that particular new form to that particular reference form, while reducing the sampling variability in the equating. However, the trade-off between bias and sampling error depended greatly on the extent to which the current equating matched the prior equatings used as collateral information with respect to the difference in difficulty between the new form and reference form.
**Purpose**

The present study was conducted as an extension of the previous study (Kim, Livingston, & Lewis, 2008). In the previous study, all the prior equatings were chosen from the same test as the targeted new forms, and were based on fairly large samples of 300 to 6,000 examinees. In the present study, the equatings used as collateral information were taken from several different tests. Also, the sample size in the equatings used as collateral information was investigated as a factor in the study.

Prior equatings to use as collateral information can be selected either from previous forms of the test to be equated or from other tests having similar characteristics (e.g., test length and test difficulty). In practice, data from the equating of previous forms may not be available for the tests where collateral information is most needed. When prior equating data are available, the equatings may not have been done in large groups of examinees. The goal of the present study was to explore the effectiveness of using collateral information on small-sample equating in situations where the prior equatings were from several different tests and were computed from samples of 100, 200, and more than 1,400 examinees.

**Method**

**Data**

The data for this study came from national administrations of licensure tests for teachers: 10 different test forms, representing 9 different subjects from a variety of content areas in education, natural science, social science, and humanities. Two different forms of one
test were included because a preliminary investigation (Kim et al., 2008) had shown the
effects of using collateral information to be quite different for those two forms.

Table 1 presents a statistical comparison of the 10 test forms and their large-group
equatings. The order of the tests in Table 1 was based on their length, which varied from 107
to 149 items. The proportion of anchor items ranged from 25% to 39%, and the correlations
between the total and anchor scores ranged from .86 to .94.

Because the 10 tests differed substantially in length, the mean scores in Table 1 are
expressed in terms of percent correct. Mean scores on the 10 new forms ranged from 58% to
76%, indicating substantial differences in difficulty among the 10 tests. As the percent-correct
anchor score means indicate, the new-form and reference-form groups for Forms 1, 2, 5, 7, 9,
and 10 were equal in ability. On three tests (Forms 3, 6, and 8), the new-form groups were
more able than the reference-form groups; on one test (Form 4), the new-form group was less
able than the reference-form group. The last row in Table 1 presents the difference between
each new form and its reference form in difficulty, as indicated by the large-group equating.
The statistic reported is the weighted average of the difference between the equated score and
the raw score, computed at each new-form score level and weighted by the population
frequency at that level. New forms 4 and 7 were about as difficult as their reference forms.
New forms 1, 2, 5, 8, 9, and 10 were more difficult than their reference forms, whereas new
forms 3 and 6 were much easier than their reference forms.

The large-group equating of each new form to its reference form was done by the
chained linear method. Figure 1 shows the distribution of these equating adjustments at new-
form score levels of 20%, 30%, etc. The equating adjustment is the difference between the
new-form raw score and equated raw score. In Figure 1 both scores are expressed in percent-
correct terms. Some new forms were more difficult and others less difficult than their reference forms. These large-group equatings were used as the criterion functions for evaluating the small-sample equatings.

**Procedure**

To determine the accuracy of a small-sample equating of two test forms, it is necessary to know or to estimate accurately the results of equating those two forms in the population that the small sample represents. We created such a situation by starting with equatings based on large groups of examinees, drawing small samples from those large groups, and applying five different equating methods to the small-sample data.

The basic procedure was as follows:

1. Select pairs of test forms that have been equated in large groups of examinees.
2. Specify new-form and reference-form sample sizes for the small-sample equating methods to be investigated.
3. Draw a small sample of the specified size from the large group of examinees who took the new form and a small sample of the specified size from the large group of examinees who took the reference form.
4. Equate the new form to the reference form in the small samples, using each of the methods to be investigated. Record the difference between the results of each small-sample equating and those of the large-group equating.
5. After a specified number of replications of steps 3 and 4, summarize the results separately for each small-sample equating method.

The evaluation included 10 sets of resampling studies, each set targeting one of the 10 new forms as the form to be equated. The equatings of the remaining 9 forms were treated as
prior equatings, to be used as collateral information. Each of the 10 sets of resampling studies included 15 separate studies, each with 500 replications of the sampling and equating procedures. The 15 studies represented all possible combinations of five new-form sample sizes (10, 25, 50, 100, and 200 examinees) and three sample sizes for the prior equatings (100, 200, and all available examinees). The new-form examinees for each replication were selected from the large groups by simple random sampling without replacement, using SAS PROC SURVEYSELECT. The reference-form sample size was held constant at 100 examinees.

The small-sample equating methods (Step 4 above) were chained linear equating, chained mean equating, chained linear equating with collateral information (“EB chained linear”), chained mean equating with collateral information (“EB chained mean”), and a synthetic equating that was robust, but not Bayesian: the average of the chained linear equating and the identity, weighted equally. Because the chained linear function received a weight of 50 percent, we call this method the “CL50 synthetic method.” The EB equatings required estimates of the variance of the current equating at each score level ($\hat{\sigma}_\text{eq}^2$ in Equation 1). These estimates were computed using the delta method. (For the chained linear equating, see Kolen & Brennan, 2004, pp. 254-255; for the chained mean equating; see Braun & Holland, 1982, p. 36.)

In each replication of each resampling study, the estimate of the prior variance ($\hat{\sigma}_\text{prior}^2$ in Equation 1) was based on the 9 prior equatings used as collateral information and the small-sample equating of the targeted new form. Thus, the set of equatings used as collateral information was nearly the same for any two of the ten targeted equatings. The sample size for the 9 prior equatings was varied as a factor in the experimental design: 100, 200, or all
available examinees. When the prior equatings were to be based on samples of 100 or 200 examinees, rather than on all available examinees, the process of sampling and computing each of the 9 prior equatings was conducted anew for each replication by simple random sampling without replacement. All 9 prior equatings were performed using the chained linear method. When the prior equatings were based on all available examinees, they could not vary over the 500 replications. In this case, only the current equating (which was also included in the estimation of the prior variance) differed in each replication. Table 2 shows the summary of the factors manipulated in our design.

**Evaluation**

The evaluation of the accuracy of the small sample equatings was based on the root-mean-squared error (RMSE) over repeated sampling, where “error” is defined as the difference between the equated scores produced by the small-sample equating and the large-group criterion equating. The RMSE can be decomposed into two orthogonal components: (1) the deviation of the individual small-sample results from their average value, i.e., the standard error of equating (SEE), and (2) the deviation of this average small-sample value from the large-group value, i.e., the statistical bias of the procedure. Note that the term “bias” in this case refers to the process of estimating the equating transformation for a specific pair of forms, over repeated sampling of examinees. This bias does not generalize to other pairs of forms. A particular equating method can be biased in one direction for one pair of forms and in the opposite direction for another pair of forms. Therefore, this bias is not the type that tends to accumulate over chains of equatings of several forms of a test. Equations 3 to 5 are the formulas for bias, error, and RMSE at a given new-form raw-score value (indicated by the subscript $i$). In these equations, $j$ indexes the replications of the procedure, $J$ is the total
number of replications (500), \( \hat{e}_j(x_i) \) is the equated score for raw score \( x_i \) in the \( j \)th replication of the procedure, \( e(x_i) \) is the equated score for raw score \( x_i \) in the criterion equating function, and \( d_i \) is the difference \( \hat{e}_j(x_i) - e(x_i) \).

\[
Bias_i = \bar{d}_i = \frac{\sum_j [\hat{e}_j(x_i) - e(x_i)]}{J},
\]

(3)

\[
SEE_i = s(d_i) = \sqrt{Var_j [\hat{e}_j(x_i) - e(x_i)]} = \sqrt{Var_j [\hat{e}_j(x_i)]},
\]

(4)

\[
RMSE_i = \sqrt{\bar{d}_i^2 + s^2(d_i)}.
\]

(5)

To compare the accuracy of the equating methods in the full examinee population, these statistics were averaged over the new-form raw-score distribution, weighting the “error” at each new-form raw-score value in proportion to the frequency of that raw score in the group of examinees taking the new form in the data set for the large-group equating. Using \( w_i \) to represent the proportional frequency at new-form raw score \( i \), the resulting statistics were: (1) the weighted root mean squared bias \(^1\) \( \sum_i w_i Bias_i^2 \), (2) the weighted standard error of equating \( \sum_i w_i SEE_i^2 \), and (3) the weighted RMSE \( \sqrt{\sum_i w_i RMSE_i^2} \). Both bias and error are equally serious and contribute equally to the RMSE. We present them separately to show how sample size affects the two components of the RMSE.

\(^1\) We computed the weighted root mean squared bias to prevent large positive and negative differences from cancelling out each other.
We will present the results that indicate the overall benefit of the EB method over the 10 targeted new forms, averaging the bias, equating error, and RMSE over the 10 forms. The averaging of the RMSE across different tests implies an assumption that RMSE values on different tests are comparable. We are not assuming that the percent-correct raw scores themselves are comparable across tests; clearly, they are not. We are assuming that measures of the variation in the percent-correct raw scores are at least approximately comparable across tests. Note that none of the 10 test forms is so easy or so difficult that the variation in the scores is restricted by a ceiling or floor effect.

**Results**

Figure 2 shows the RMSE for each equating method, as a function of the new-form sample size and the size of the samples in the prior equatings used as collateral information. Figures 3 and 4 show the two components of the RMSE separately: the root mean squared bias and equating error. The statistics shown in these figures were overall values, computed across 500 replications, averaged over score distributions on the new form in the full population (i.e., the large group described by the statistics in Table 1), and then averaged over the 10 targeted new forms. The dashed horizontal line in Figure 2 indicates the RMSE that would result from not equating, i.e., from assuming the new and reference forms to be equally difficult.

Figure 2 shows the RMSE for 45 combinations of equating method, size of the new-form equating sample, use of collateral information, and sample size for generating the collateral information. All but one of the 45 combinations yielded a smaller RMSE than the use of the identity – in most cases, much smaller. The single exception was chained linear
equating with no collateral information and with new-form samples of only 10 examinees. In general, the choice of method, the presence or absence of collateral information, and the size of the samples used to generate the collateral information all had only a small effect on the RMSE, in comparison to the effect of increasing the new-form sample size. The exceptions to this generalization were (1) the very large RMSE for chained linear equating with new-form samples of only 10 examinees and no collateral information and (2) the large RMSEs for the CL50 synthetic method with new-form samples of 50 or more examinees. Even with large samples for the prior equatings, using them as collateral information did not improve the accuracy of equating when the new-form sample included 25 or more examinees.

Figures 3 and 4 separate the RMSE into its two orthogonal components: bias (Figure 3) and random error (Figure 4). “Bias” in this case refers to the difference between the average over replications of the small-sample equating results for a particular pair of forms and the equated score determined in the large-group criterion equating. The dashed horizontal line for the identity does not appear in Figure 4 because this procedure reduces the random error to zero (at the cost of a substantial bias). The EB chained linear and EB chained mean methods yielded much greater bias than their non-EB counterparts, but substantially less equating error, particularly at the smaller sample sizes.

There were some differences between the chained linear and chained mean methods as a function of bias and equating error. The chained mean method displayed the smallest RMSE for samples as small as 50, whereas the chained linear method displayed the smallest RMSE for samples larger than 50; the same trend emerged for their EB counterparts. Increasing the sample size for the prior equatings from 100 to 200 has some effect on the accuracy of the EB equating; any further increase has essentially no effect. In this context, the reduction in
random error did not compensate for the increased bias associated with the EB method. The
CL50 synthetic method did not perform as well as chained mean equating for any new-form
sample size, although it performed nearly as well for samples of only 10 examinees. The
synthetic CL50 method performed quite poorly with the sample of 50 or more examinees
because the weights assigned to the current equating (chained linear in this case) did not
depend on sample size. As shown in Figure 3, the magnitude of bias was constantly large
across all sample sizes.

The purpose of this series of resampling studies was to determine whether collateral
information from different tests would enhance the accuracy of equating in small samples of
examinees. Figures 5 to 7 answer this question separately for each pair of forms to be
equated. Each data point in this graph displayed using numbers represents one of the 10 test
forms to be equated. Specifically, numbers 1 to 9 correspond to Forms 1 to 9 in Table 1,
respectively, and number 0 corresponds to Form 10. The horizontal position of the point (x-
value) indicates the weighted RMSE for the equating of that form without collateral
information. The vertical position of the point (y-value) indicates the weighted RMSE for the
equating using collateral information. Data points to the lower right of the diagonal line
indicate new forms for which the collateral information improved equating accuracy.

Each of the weighted RMSEs shown in Figures 5 to 7 was computed across 500
replications and then averaged over the score distributions on the new form in the full
population. Figures 5 to 7 show the results using prior equatings based on samples of 200
examinees. The results using prior equatings based on samples of 100 examinees and those
using prior equatings based on all available data were similar to those shown in Figures 5 to 7.
The results for new-form samples of 200 examinees are not shown; they were essentially the same as those for new-form samples of 100 examinees.

Figure 5 compares the EB and non-EB versions of chained linear equating. The first plot in Figure 5, for new-form equating samples of only 10 examinees, shows that the use of collateral information produced a large improvement in the accuracy of chained linear equating for 4 of the 10 new forms, a moderate improvement for 1 form, a small improvement for 3 forms, no improvement for 1 form, and much less accurate equating for 1 form. The second plot, for new-form equating samples of 25 examinees, shows that collateral information improved equating accuracy for 4 forms, had a very small effect on 4 forms, made equating somewhat less accurate for 1 form, and made it much less accurate for 1 form.

The numbers presented in the last row of Table 1, which show how each new form differed from its reference form in difficulty, explain these results. Of the 10 new forms, 2 Forms 4 and 7 were similar to their reference forms in difficulty (differing by less than 1 percentage point); Forms 1, 2, 5, 8, 9, and 10 were more difficult than their reference forms, and Forms 3 and 6 were easier. These two easier new forms account for the highest two data points in each plot of Figure 5, the forms for which collateral information tended to make chained linear equating less accurate. New Form 3, which yielded the highest data point was nearly four percentage points easier than its reference form.

Figure 6 shows the same comparison as Figure 5, but for the chained mean equating method. The first plot of Figure 6 presents the results for new-form samples of 10 examinees. It shows that the use of collateral information yielded a substantial improvement for 5 of the 10 new forms, but almost no improvement for the other 4 forms. This plot shows one data point well above the diagonal line, representing New Form 3, the new form that was much
easier than its reference form. This was the one new form for which the use of collateral information made chained mean equating clearly less accurate. In the second plot, for new-form samples of 25 examinees, again the data point for this new form is substantially higher than the others. In general, the second, third, and fourth plots of Figure 6 show that for this group of 10 new forms, the use of collateral information did not substantially improve the accuracy of chained mean equating with new-form samples of 25 or more examinees.

Figure 7 compares the accuracy of chained linear equating (without collateral information) with that of the CL50 synthetic method (the equally-weighted average of the chained linear equating and the identity). The first plot, for new-form samples of 10 examinees, shows this use of the identity making a substantial improvement over the chained linear equating for 6 of the 10 new forms, but no improvement for the other four new forms (Forms 2, 3, 8, and 10) that differed substantially from their reference forms in difficulty. The second plot shows that with new-form samples of 25 examinees, adding in the identity actually made the chained linear equating less accurate for those four new forms. The lowest two data points in the fourth plot corresponds to New Forms 4 and 7, which were very similar to their reference forms in difficulty (differing by less than 1 percentage point).

**Discussion**

The smaller the samples available for equating, the greater the value of collateral information from the equating of other test forms. But often, small-volume tests have few prior forms that have been equated – or none at all. The applicability of EB equating methods would be extremely limited if the prior equatings used as collateral information were restricted to forms of the same test as the targeted new form. The present study investigated
the effectiveness of using collateral information from prior equatings of other tests. In the present study, the collateral information came from other tests and was derived from samples of three different sizes: small \((N = 100)\), medium \((N = 200)\), and large \((N > 1,400)\).

RMSE statistics, computed over sampling replications and averaged across 10 test forms to be equated (representing 9 different tests), indicated that the EB procedure improved the accuracy of chained linear equating with new-form samples of 10 examinees, but not with new-form samples of 25 or more examinees. The EB procedure did not substantially improve the accuracy of mean equating even with new-form sample of only 10 examinees. (In all cases, the reference-form sample included 100 examinees.) Averaging the RMSE over the 10 test forms, however, conceals an important finding. For 5 of the 10 new forms, the EB procedure resulted in a substantial improvement in accuracy when the new-form sample was small, but for one new form, the EB procedure made the equating substantially less accurate. This form differed from the other 9 forms in the study by being much easier than the reference form to which it was being equated. Most of the other 9 forms were either more difficult than their reference forms or of similar difficulty. Varying the size of the equating samples used to provide the collateral information had only a small effect.

What do the results of the study tell us about the usefulness of collateral information from tests other than the test to be equated? The most important factor in determining whether collateral information will improve the accuracy of small-sample equating appears to be way in which new forms of that test tend to differ in difficulty from the reference forms they are equated to. If the form being equated follows the pattern established by the forms used as collateral information, the collateral information will improve the accuracy of equating. If it does not, the collateral information will not help and may even make the equating less
accurate. Either pattern of results can occur, no matter whether the collateral information comes from other forms of the same test or from forms of other tests. (See Kim, et al., 2008, for two examples with collateral information drawn from previous forms of the test being equated.)

Some experts have suggested simply using the identity function as the equating transformation, when the samples available are smaller than 100 examinees per form (see Kolen & Brennan, 2004, p. 100). In the present study, the use of the identity as the equating transformation yielded a much larger RMSE than any of the methods investigated, at all new-form sample sizes investigated, with only one exception: the chained linear method with new-form samples of only 10 examinees and no collateral information. Even with only 10 examinees in the new form equating sample, the EB chained linear equating, both EB and non-EB versions of mean equating, and the CL50 synthetic method all yielded results much more accurate than the results of using the identity as the equating transformation.

There may be a much better way to use the identity. The EB procedure investigated in this study involves an implicit assumption that, in the absence of any information about the difficulty of a specific new form, the best assumption is that it will differ from its reference form in the same way as did the new forms in the equatings used as collateral information. That assumption may not be a good one, particularly if the number of equatings used as collateral information is not large. An alternative would be to use, as a prior estimate, the mean that would result from equating all possible pairs of forms in both directions (i.e., equating Form X to Form Y and also equating Form Y to Form X). That mean would be the identity. The role of the collateral information would then be limited to estimating the variance of the prior equatings. This variance would be used as the variance of the prior
estimate in the EB formula (even though the variance of the identity is actually zero), to
determine the relative weight of the identity and the small-sample results.

The resulting procedure would be like the CL50 synthetic method investigated in this study, but with weights determined empirically instead of being specified a priori. The results of the CL50 synthetic method with new-form samples of only 10 examinees make this method appear promising; for 6 of the 10 test forms this procedure worked substantially better than chained linear equating, and for the other 4 forms it worked as well. The main problem with the CL50 synthetic method is that the weights it assigns to the current equating and the identity do not depend on sample size. As the sample size increases, the observed equating becomes a better estimate of the population equating function and should receive greater weight. The EB formula provides a way to determine weights that have this property.

Even with this alternative method, the question remains: Should equatings of other tests be used in determining the variance associated with the prior estimate? If so, how can a user decide whether to use equatings of a particular test for this purpose? The general principle to follow would be to limit the collateral information to tests that are similar to the test to be equated – similar with respect to the factors that affect the extent to which forms of the same test vary in difficulty. These factors would include the approximate number of items, the response mode (multiple-choice, short-answer, etc.), and – possibly more than any other factor – the availability of pretest information on the items.

For some tests, all new items are pretested on large, representative samples of the examinee population. The test-developers assembling each new form use the pretest information to select items that meet a set of difficulty specifications. New forms of these tests tend to be similar to their reference forms in difficulty, and the equating adjustments tend
to be small. For some other tests, the items cannot be pretested. The test-developers assembling each new form have only their intuition and experience to guide them in anticipating the difficulty of the items. New forms, assembled without benefit of pretest data, often differ considerably from their reference forms in difficulty, and size of the equating adjustments tends to vary from one form to another.

One finding that was somewhat surprising (to us) was the strong performance of mean equating – without collateral information – in new-form samples of 25 and 50 examinees. The chained mean equating method consistently yielded smaller RMSEs than the other methods for those small samples. Given that specifying collateral information is a demanding task, if the improvements it produces are very small, the method of choice for equating in extremely small samples could be chained mean equating without collateral information.
References


<table>
<thead>
<tr>
<th>Number of items</th>
<th>Group</th>
<th>Form 1</th>
<th>Form 2</th>
<th>Form 3</th>
<th>Form 4</th>
<th>Form 5</th>
<th>Form 6</th>
<th>Form 7</th>
<th>Form 8</th>
<th>Form 9</th>
<th>Form 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>148</td>
<td>130</td>
<td>120</td>
<td>119</td>
<td>119</td>
<td>118</td>
<td>115</td>
<td>108</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>149</td>
<td>130</td>
<td>118</td>
<td>119</td>
<td>118</td>
<td>120</td>
<td>116</td>
<td>108</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>Number of anchor items</td>
<td>New</td>
<td>40</td>
<td>42</td>
<td>39</td>
<td>30</td>
<td>37</td>
<td>35</td>
<td>30</td>
<td>36</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>40</td>
<td>42</td>
<td>39</td>
<td>30</td>
<td>37</td>
<td>35</td>
<td>30</td>
<td>36</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>% of anchor items</td>
<td>New</td>
<td>27%</td>
<td>32%</td>
<td>33%</td>
<td>25%</td>
<td>31%</td>
<td>29%</td>
<td>25%</td>
<td>31%</td>
<td>32%</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>27%</td>
<td>32%</td>
<td>33%</td>
<td>25%</td>
<td>31%</td>
<td>30%</td>
<td>25%</td>
<td>31%</td>
<td>32%</td>
<td>39%</td>
</tr>
<tr>
<td>Number of examinees</td>
<td>New</td>
<td>2,731</td>
<td>5,571</td>
<td>3,582</td>
<td>1,692</td>
<td>7,158</td>
<td>6,259</td>
<td>4,605</td>
<td>3,203</td>
<td>6,019</td>
<td>6,426</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>1,425</td>
<td>6,205</td>
<td>4,259</td>
<td>3,010</td>
<td>9,859</td>
<td>3,713</td>
<td>7,683</td>
<td>3,383</td>
<td>6,386</td>
<td>6,489</td>
</tr>
<tr>
<td>Total score mean (%)</td>
<td>New</td>
<td>64%</td>
<td>58%</td>
<td>74%</td>
<td>62%</td>
<td>66%</td>
<td>65%</td>
<td>76%</td>
<td>67%</td>
<td>73%</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>67%</td>
<td>62%</td>
<td>68%</td>
<td>64%</td>
<td>67%</td>
<td>61%</td>
<td>75%</td>
<td>68%</td>
<td>74%</td>
<td>72%</td>
</tr>
<tr>
<td>Anchor score mean (%)</td>
<td>New</td>
<td>66%</td>
<td>61%</td>
<td>75%</td>
<td>62%</td>
<td>69%</td>
<td>71%</td>
<td>74%</td>
<td>69%</td>
<td>76%</td>
<td>73%</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>66%</td>
<td>61%</td>
<td>72%</td>
<td>64%</td>
<td>69%</td>
<td>68%</td>
<td>74%</td>
<td>66%</td>
<td>76%</td>
<td>73%</td>
</tr>
<tr>
<td>Correlation of</td>
<td>New</td>
<td>.93</td>
<td>.93</td>
<td>.92</td>
<td>.92</td>
<td>.93</td>
<td>.87</td>
<td>.90</td>
<td>.91</td>
<td>.88</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>Ref</td>
<td>.93</td>
<td>.94</td>
<td>.92</td>
<td>.93</td>
<td>.93</td>
<td>.86</td>
<td>.89</td>
<td>.90</td>
<td>.88</td>
<td>.92</td>
</tr>
<tr>
<td>Difference in difficulty</td>
<td></td>
<td>+2.73%</td>
<td>+3.73%</td>
<td>-3.92%</td>
<td>+0.02%</td>
<td>+1.57%</td>
<td>-2.09%</td>
<td>-0.56%</td>
<td>+3.14%</td>
<td>+1.27%</td>
<td>+3.88%</td>
</tr>
</tbody>
</table>

Note: Regarding the difference in difficulty, positive value indicates new form more difficult.
TABLE 2

Factors in the Design of the Resampling Studies

<table>
<thead>
<tr>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of prior equatings used as collateral information</td>
</tr>
<tr>
<td>2. Sample size for prior equatings</td>
</tr>
<tr>
<td>3. New-form sample size for small-sample equatings</td>
</tr>
<tr>
<td>4. Reference-form sample size for small-sample equatings</td>
</tr>
<tr>
<td>5. Number of replications of sampling and equating</td>
</tr>
<tr>
<td>6. Equating method for large-group criterion equatings</td>
</tr>
<tr>
<td>7. Equating method for prior equatings</td>
</tr>
<tr>
<td>8. Equating methods for small-sample equatings</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
</tr>
<tr>
<td>100, 200, or all available examinees (1,400+)</td>
</tr>
<tr>
<td>10, 25, 50, 100, or 200</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>chained linear</td>
</tr>
<tr>
<td>chained linear</td>
</tr>
<tr>
<td>chained linear, EB chained linear, chained mean, EB chained mean, CL50 synthetic</td>
</tr>
</tbody>
</table>
Figure 1. Difference plots between percent correct raw scores and percent correct equated raw scores for the 10 targeted new test forms.

Note. The numbers 1 to 9 in this figure correspond to New Forms 1 to 9 in Table 1, respectively. For simplicity, New Form 10 was displayed using the number 0.
FIGURE 2. RMSE over 500 replications, averaged over 10 tests.

Note. The dashed line indicates the RMSE value for the identity (i.e., assuming the new form and reference form to be of equal difficulty).
FIGURE 3. Root mean squared bias over 500 replications, averaged over 10 tests.

Note. The dashed line indicates the root mean square bias value for the identity (i.e., assuming the new form and reference form to be of equal difficulty).
FIGURE 4. Equating error over 500 replications, averaged over 10 tests.
FIGURE 5. Root mean squared error for chained linear equating with and without collateral information based on samples of 200 examinees.

Note. The numbers 1 to 9 represent New Forms 1 to 9, respectively, and the number 0 represents New Form 10.
FIGURE 6. Root mean squared error for chained mean equating with and without collateral information based on samples of 200 examinees.

Note. The numbers 1 to 9 represent New Forms 1 to 9, respectively, and number 0 represents New Form 10.
FIGURE 7. Root mean squared error for the chained linear and CL50 synthetic methods.

Note. The numbers 1 to 9 represent New Forms 1 to 9, respectively, and the number 0 represents New Form 10.