Chapter 7
New Approaches to Equating With Small Samples

Samuel A. Livingston and Sooyeon Kim

7.1 Overview

The purpose of this chapter is to introduce the reader to some recent innovations intended to solve the problem of equating test scores on the basis of data from small numbers of test takers. We begin with a brief description of the problem and of the techniques that psychometricians now use in attempting to deal with it. We then describe three new approaches to the problem, each dealing with a different stage of the equating process: (1) data collection, (2) estimating the equating relationship from the data collected, and (3) using collateral information to improve the estimate. We begin with Stage 2, describing a new method of estimating the equating transformation from small-sample data. We also describe the type of research studies we are using to evaluate the effectiveness of this new method. Then we move to Stage 3, describing some procedures for using collateral information from other equatings to improve the accuracy of an equating based on small-sample data. Finally, we turn to Stage 1, describing a new data collection plan in which the new form is introduced in a series of stages rather than all at once.

7.2 The Problem

Equating test scores is a statistical procedure, and its results, like those of most other statistical procedures, are subject to sampling variability. The smaller the samples of test takers from which the equating is computed, the more the equating results are likely to deviate from what they would be in a different pair of samples—or in the population that the samples represent. For tests equated through randomly

S.A. Livingston (✉) and S. Kim
Educational Testing Service, Rosedale Rd, Princeton, New Jersey 08541, USA
e-mail: slivingston@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_7,

equivalent groups, with no anchor, the important number is the total number of test takers available for the equating administration—possibly 200, 100, or even fewer. For tests equated through common items, the problem becomes particularly acute when very few test takers take the new test form at its first administration—possibly 30, 20, or even fewer. As the psychometricians responsible for the equating of the scores, we need to find a raw-to-scale score conversion that will make the scores of the test takers who take the new form comparable to the scores of test takers who took other forms of the test. We cannot make the problem go away by simply claiming that the test takers whose scores we can observe in time for the equating are the entire target population for the equating of those two test forms. In many cases, a test form taken at first by only a few test takers will be administered later to many others. We can accumulate data over two or more administrations of the new form and re-equate the scores, but the scores of the first group of test takers will already have been reported.

Even if the new form will not be administered again, the problem remains. The important principle is that an individual test taker's reported score should not depend heavily (ideally, not at all) on the particular group of test takers whose data are used to equate the scores on the form that the test taker happened to take.[1] We need to determine an equating relationship that will generalize to other groups that may differ in ability. What we really want is a good estimate of the equipercentile equating relationship in the population of potential test takers—not simply an equating of means and standard deviations on the two forms, but an equating of the full score distributions.

7.3 Current Practice

One way to improve the accuracy of equipercentile equating in small samples of test takers is to presmooth the score distributions (see Livingston, 1993). However, if the samples of test takers are quite small, this technique may not reduce the sampling variability in the equating to an acceptable level.

Another procedure that has been recommended is to establish a minimum sample size for equating. If the available samples meet the sample size requirement, equate the scores; if samples of the specified size are not available, assume the new form and reference form to be of equal difficulty throughout the score range (see Kolen & Brennan, 1995, p. 272; Kolen & Brennan, 2004, pp. 289-290). We believe there are better ways to deal with the problem.

Possibly the most common approach to estimating a relationship in a population on the basis of small-sample data is to use a strong model. Strong models require

[1]This principle is the basis for the requirement of *population invariance* (see, e.g., Dorans, Moses, & Eignor, Chapter 2 of this volume). In the case of equating with small samples, a greater problem is that the samples of test takers may not adequately represent any population.

only a small number of parameters to be estimated from the data, in effect substituting assumptions for data. In test score equating, the strong model most commonly used is the linear equating model. Its basic assumption is that in the target population, the distributions of scores on the new form (to be equated) and on the reference form (to which it is being equated) differ only in their means and standard deviations. An even stronger model is that of *mean equating*, a linear equating model that assumes that the score distributions on the new form and reference form in the target population differ only in their means (see Kolen & Brennan, 1995, pp. 29-30; Kolen & Brennan, 2004, pp. 30-31). Both of these models constrain the equating relationship to be linear. However, when test forms differ in difficulty, the equating relationship between them typically is not linear. If the difficulty difference is substantial, the relationship is not even approximately linear. A harder form and an easier form, administered to the same group of test takers, will tend to produce differently skewed distributions. The stronger test takers' scores will tend to vary more on the harder form than on the easier form; the weaker test takers' scores will tend to vary more on the easier form than on the harder form. Consequently, the slope of the equating transformation will not be the same for the weaker test takers as for the stronger test takers. A linear transformation, with its constant slope, cannot adequately model the equating relationship.

7.4 Circle-Arc Equating

Circle-arc equating is a strong model that does *not* assume the equating relationship to be linear. It is based on an idea from Divgi (1987). Divgi's idea was to constrain the equating curve to pass through two prespecified end points and an empirically determined middle point. Although the circle-arc model is different from Divgi's, it also constrains the equating curve to pass through two prespecified end points and an empirically determined middle point. In circle-arc equating, the lower end point corresponds to the lowest meaningful score on each form. On a multiple-choice test scored by counting the number of correct answers, the lowest meaningful score would typically be the chance score—the expected score for a test taker who responds at random (e.g., without reading the questions). The upper end point corresponds to the maximum possible score on each form. The middle point is determined by equating at a single point in the middle of the score distribution.

7.4.1 The Circle-Arc Method

The circle-arc equating method requires only one point on the equating curve to be estimated from the small-sample data. The first step of the method is to determine that point. The user of the method can choose the x-value at which to make the estimate, and that x-value need not be a score that actually can be obtained on the

test. The part of the score scale where the equated scores can be estimated most accurately, particularly in small samples, is the middle of the distribution. If the equating is a direct equating (e.g., an equivalent-groups equating), the middle point can be determined by equating the mean score on the new form directly to the mean score on the reference form. If the equating is through an anchor score (e.g., a common-item equating), the middle point can be determined by equating at the mean score of the smaller group of test takers. Typically, the smaller group will be the group taking the new form.

If the middle point happens to lie on the line connecting the end points, that line is the estimated equating curve. If not, the next step is to use the end points and the middle point to determine the equating curve. There are two versions of circle-arc equating, and they differ in the way they fit a curve to these three points. We call one version *symmetric circle-arc equating* and the other *simplified circle-arc equating*. Symmetric circle-arc equating is actually simpler conceptually, but its formulas are a bit cumbersome. Simplified circle-arc equating uses a slightly more complicated model in order to simplify the formulas. In the research studies we have done (which we describe later in this chapter), the two versions of the circle-arc method have produced about equally accurate results. The formulas for both versions appear in the Appendix to this chapter. Both versions are described in Livingston and Kim (2008); only the simplified version is included in Livingston and Kim (2009), but the formulas for the symmetric version are given in Livingston and Kim (2010).

Both methods are applications of the geometrical fact that if three points do not lie on a straight line, they uniquely determine a circle. Symmetric circle-arc equating fits a circle arc directly to the three data points. Simplified circle-arc equating transforms the three data points by decomposing the equating function into a linear component and a curvilinear component (an idea borrowed from von Davier, Holland, & Thayer, 2004b, pp. 11–13). The linear component is the line connecting the two end points. The curvilinear component is the vertical deviation of the equating curve from that line. It is estimated by fitting a circle arc to the three transformed data points.

Figure 7.1 illustrates the simplified circle-arc procedure. The horizontal axis represents the score on the new form, that is, the test form to be equated. The vertical axis represents the corresponding score on the reference form. The two prespecified end points and the empirically determined middle point are indicated by the three small circles. The line connecting the two end points is the linear component of the estimated equating curve. The three data points are transformed by subtracting the y-value of that line, which we call $L(x)$. The three transformed points are indicated by the squares at the bottom of Figure 7.1. Both end points are transformed onto the horizontal axis. The middle point is transformed to a point above or below the horizontal axis—above it if the new form is harder than the reference form, and below it if the new form is easier than the reference form. In the example illustrated by Figure 7.1, the new form is harder than the reference form. Consequently, the original middle point is above the line $L(x)$, and the transformed middle point is above the horizontal axis.

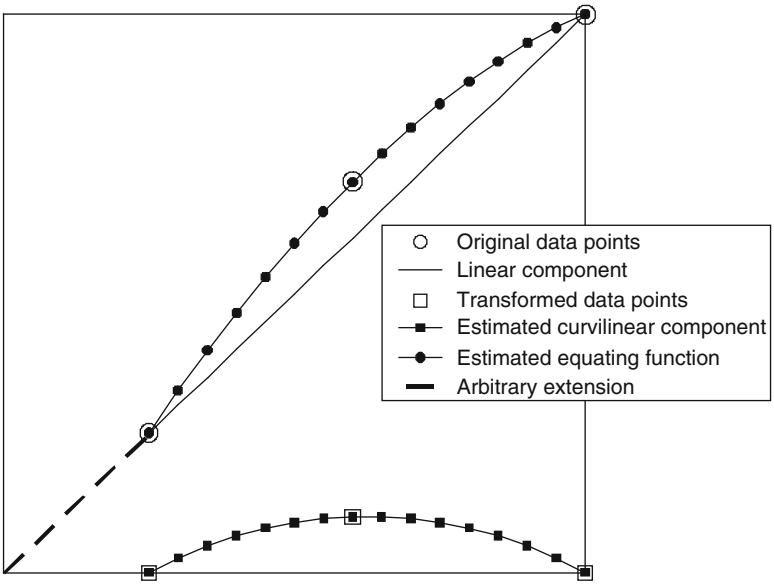

Fig. 7.1 Illustration of the simplified circle-arc equating method

The three transformed data points uniquely determine an arc of a circle. The arc connecting the three transformed data points is shown at the bottom of Figure 7.1. This arc serves as an estimate of the curvilinear component of the equating function. For each possible raw score on the new form, there is a corresponding point on the arc. The next step is to add the linear component back in, by adding the height of the line $L(x)$ to the height of the arc. The three original data points are retransformed back to their original positions, and the full arc is transformed into an estimate of the equipercentile equating function, shown in the upper portion of the figure. The last step is to extend the equating transformation below the lower end point, by connecting that point linearly to the point corresponding to the minimum possible score on each form. This line is arbitrary—hardly a problem, since the lower end point of the curve corresponds to the lowest meaningful score on each form.

Purists may object that simplified circle-arc equating is not truly an equating method, because it is not symmetric in its treatment of the new form and reference form. Indeed, it is not symmetric, but we are not suggesting either circle-arc method as a way to define the equating relationship in the population. We are suggesting them as tools for estimating the equipercentile equating relationship in the population. The equipercentile equating relationship is symmetric, but the best available procedure for estimating it from small-sample data may not be symmetric.

Because the end points of the estimated equating curve are constrained, the estimate produced by either circle-arc method has a sampling variance of zero at the end points. The conditional standard error of equating (CSEE) at those points is zero. At the middle point, the CSEE depends on the equating method used to determine the y value for that point. The CSEE at that point can be estimated by

whatever procedure is appropriate for that method. At any other point, the CSEE can be approximated by a simple proportion; the CSEE values at any two points are approximately proportional to their deviations from the line connecting the two end points. If (x_2, y_2) is the middle point, for which an estimate of the CSEE is available, and (x_j, y_j) is the point for which the CSEE is to be estimated, then

$$\frac{CSEE(y_j)}{CSEE(y_2)} \doteq \left(\frac{y_j - L(x_j)}{y_2 - L(x_2)}\right) \tag{7.1}$$

7.4.2 Resampling Studies

We have been conducting resampling studies to evaluate the accuracy of equating in small samples by several methods, including the two circle-arc methods. We have investigated both common-item equating and random-groups equating, using somewhat larger sample sizes in the random-groups design. The basic procedure for a random-groups design is as follows (Livingston & Kim, 2010):

Choose an existing test form of approximately 100 or more items, a form that has been administered to several thousand test takers. Consider those test takers as the target population for equating. Divide the test form into two nonoverlapping subforms, parallel in content but unequal in difficulty. Designate one subform as the new form and the other as the reference form for equating. Compute the direct equipercentile equating of the new form to the reference form in the full target population; this equating is the criterion equating for the resampling study. To evaluate the small-sample equating methods, for a given sample size, perform several hundred replications of this procedure:

1. Draw a pair of nonoverlapping samples of test takers, sampling from the full population.
2. Compute the distribution of scores on the new form in one sample of test takers and on the reference form in the other sample.
3. Use those score distributions to equate the new form to the reference form, by all the small-sample methods to be compared.
4. At each new-form raw-score level, record the difference between the results of each small-sample equating method and the criterion equating.

Summarize the results for each small-sample equating method, summarizing over the several hundred replications by computing, for each raw score on the new form, the root-mean-square deviation (RMSD) of the sample equatings from the population equating.

We repeated this procedure with six operational test forms, each from a different test. To average the results over the six test forms, we expressed the RMSD values in terms of the standard deviation of the scores on the reference form in the full population. We then conditioned on percentiles of the distribution of scores on the

new form in the population and took the root mean square, over the six test forms, of the RMSD values at those percentiles. The result is a set of curves, one for each small-sample equating method. Figure 7.2 shows the resulting curves for the two circle-arc methods, for three other equating methods, and for the identity, with samples of 200 test takers for each form (a small sample size for random-groups equating). In the middle of the distribution, all the equating methods performed about equally well. At the low end of the distribution, the two circle-arc methods and mean equating were more accurate than linear or equipercentile equating. At the high end of the distribution, the two circle-arc methods were much more accurate than the other methods at estimating the equipercentile equating in the population.

The resampling procedure for common-item equating (Kim & Livingston, 2010) was a bit more complicated. Again, we used data from test forms taken by several thousand examinees, but in this case we selected forms that had been given on two separate occasions to populations that differed somewhat in ability. As in the random-groups studies, we used the items in the full test form as an item pool to construct subforms to be equated, but in this case the subforms included a set of items in common, for use as an anchor in the small-sample equatings. The criterion equating was the direct equipercentile equating of the two subforms in the combined population of test takers taking the full test form. In the resampling studies, instead of selecting both new-form and reference-form samples from the same population of test takers, we designated the test takers from one testing occasion as the new-form population and those from the other testing occasion as the reference-form population. The sample sizes we investigated were smaller than those in the random-groups studies. We also specified the reference-form sample to

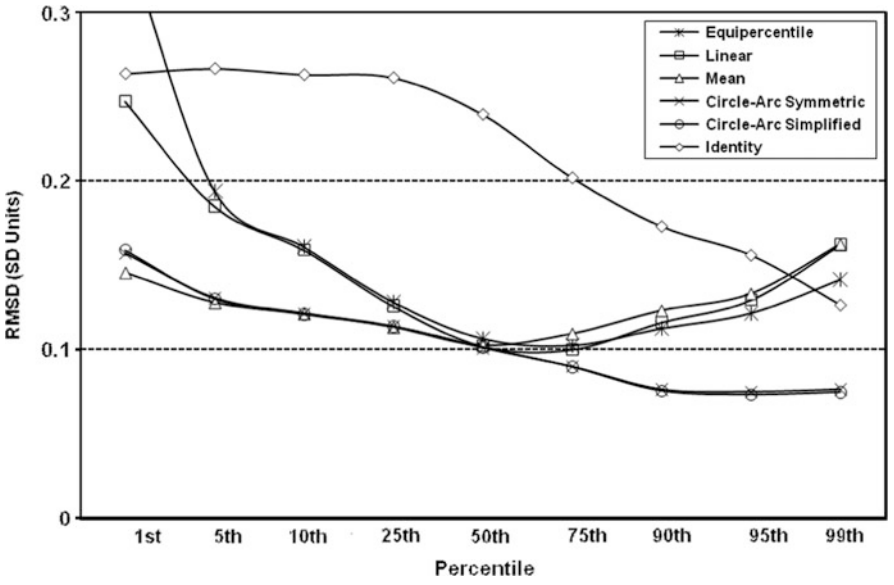

Fig. 7.2 Example of resampling study results

be 3 times as large as the new-form sample, to reflect the usual small-sample common-item equating situation, in which reference-form data are accumulated over two or more testing occasions. The results of the common-item equating studies in small samples (50 or smaller for the new form, 150 or smaller for the reference form) were generally similar to those of the random-groups studies, with one difference: At the low end of the score distribution, mean equating was clearly the most accurate method, even more accurate than the circle-arc methods (which were the most accurate in the upper portion of the score distribution).

7.5 Equating With Collateral Information

When we psychometricians equate test scores on the basis of data from very small groups of test takers, we are trying to estimate the equating relationship we would find if we had data from very large groups of test takers. One way to improve an estimation process, especially when the data come from a small sample, is to incorporate collateral information (see, e.g., Efron & Morris, 1977). Collateral information for equating test scores is often available from equatings of other forms of the test we want to equate and of other tests.

7.5.1 Empirical Bayes Estimation

The idea of using collateral information suggests an empirical Bayes approach. The basic premise is that the current, small-sample equating can be regarded as a single observation randomly sampled from a large domain of possible equatings, each with its own new form, reference form, and samples of test takers. For a given pair of test forms, there is a "true" equating function: the function that would result from averaging over all the equatings of that particular pair of test forms with different samples of test takers. All the equatings in the domain are assumed to provide information that may be relevant for estimating this true equating function. Therefore, how broadly to define the domain is an important question. (We discuss it briefly under Section 7.5.2, Problems and Limitations, below.)

Developing a Bayesian estimate for an equating function turns out to be a complex problem. A practical way to simplify the problem is to estimate the equating transformation *one point at a time*. For each possible raw score on the new form, there is a single equated score on the reference form to be estimated. Estimating a single equated score is a much more manageable task than estimating the equating transformation all at once.[2]

[2]For an alternative approach based on modeling the discrete bivariate distribution of scores on the two forms to be equated, see Karabatsos and Walker (Chapter 11 of this volume).

For a given raw score x on the new form, the empirical Bayes estimate of the corresponding equated score y on the reference form is a weighted average of a "current" estimate and a "prior" estimate, which we will call $y_{current}$ and y_{prior}. The current estimate is the equated score implied by the small-sample equating. The prior estimate is the average of the equating results in all the equatings used as collateral information, with the addition of the current equating. The current equating is included in the prior estimate because the prior estimate is intended to represent a domain of possible equatings, and the current equating is a member of the domain.

The Bayesian formula for combining the current and prior estimates is

$$\begin{aligned} \ddot{y}_{EB} &= \frac{\frac{1}{\text{var}(y_{current})} y_{current} + \frac{1}{\text{var}(y_{prior})} y_{prior}}{\frac{1}{\text{var}(y_{current})} + \frac{1}{\text{var}(y_{prior})}} \\ &= \frac{\left[\text{var}(y_{prior})\right] y_{current} + \left[\text{var}(y_{current})\right] y_{prior}}{\text{var}(y_{prior}) + \text{var}(y_{current})} \end{aligned} \tag{7.2}$$

The smaller the samples of test takers in the current equating, the more unstable the results are likely to be, resulting in a larger variance for $y_{current}$ and a smaller weight for it in the empirical Bayes estimate. On the other hand, the fewer equatings that contribute to the prior estimate and the more those equatings differ, the larger the variance of y_{prior} and the smaller the weight for it in the empirical Bayes estimate.[3]

7.5.2 Problems and Limitations

One feature of the real world of testing that complicates this Bayesian procedure is that test forms differ in length. Even alternate forms of the same test sometimes differ in length, because of the exclusion of one or more items from the scoring. One solution to this "apples and oranges" problem is to convert the scores on all the forms, in the current equating and in all the prior equatings, to percentages. This tactic creates a common metric and makes it possible to use interpolation to determine corresponding scores on forms that do not have exactly the same number of items.

Another difficulty in implementing this approach is that the empirical Bayes formula requires, at each new-form raw-score value, an estimate of the sampling variance of the current equating (i.e., the square of the CSEE). An estimate computed from the small-sample data in the current equating is likely to be

[3]We thank Charles Lewis for his help in working out the details of this procedure. A paper by Livingston and Lewis (2009) contains a more complete description and explanation of the procedure.

inaccurate. We have been investigating the possibility of using data from many equatings to develop estimates of this variance, as a function of sample size, test length, and approximate test difficulty.

Yet another difficulty in implementing this approach is that of deciding what equatings to include as collateral information. Should the collateral information include equatings of other tests? If so, how different from the test to be equated can another test be and still provide useful information? This question is an empirical question, and the answers may differ for different kinds of tests. We have been doing some research (consisting of resampling studies), and the results indicate that the most important factor is, not surprisingly, the way in which new forms of the test differ in difficulty from the forms to which they are equated (Kim, Livingston, & Lewis, 2008, 2009).

To see what can go wrong, consider a situation in which the new form to be equated is easier than the reference form, but in all the prior equatings used as collateral information, the new form was harder than the reference form. In this case, the collateral information will pull the equated score toward a value that is typical for the domain but wrong for the current equating. The same problem will occur if the new form to be equated is much harder or much easier than the reference form but in all the prior equatings the new form and reference form were very similar in difficulty.

7.5.3 A Simpler Approach to Using Collateral Information

The empirical Bayes procedure described above is highly sensitive to the choice of equatings used as collateral information. In addition, the calculations involved are laborious. However, there is another procedure, based on similar reasoning, that does not depend as heavily on the choice of collateral information and is also simpler to use operationally.

Consider the prior estimate of the equated score at each score level. The Bayesian procedure described above uses the mean of a group of equatings. Its results depend heavily on the choice of those equatings. Now suppose the domain of equatings that provide collateral information includes two equatings for each pair of test forms—equating form X to form Y and equating form Y to form X. In that case, averaging over all the equatings in the domain will yield a result very close to the identity ($Y = X$). Instead of using the mean of a specified set of equatings, we can simply use the identity as the prior estimate toward which the small-sample equating results will be pulled.

Using the identity is not a new idea. Some writers have advocated using the identity as the equating function whenever the size of the samples available falls below a specified threshold—one that depends on the extent to which test forms are expected to differ in difficulty (Kolen & Brennan, 1995, p. 272; Kolen & Brennan, 2004, pp. 289-290; Skaggs, 2005, p. 309). We think there is a better way to take sample size into account. Instead of making an “either–or” decision, compute a

weighted average of the small-sample equating and the identity. For a given raw-score x, if $y_{obs}(x)$ represents the equated score observed in the small-sample equating, the adjustment is simply

$$y_{adj}(x) = w[y_{obs}(x)] + (1 - w)x\,, \tag{7.3}$$

where w is a number between 0 and 1.

The practical question for implementing this procedure is how to choose a value for w. Kim, von Davier, and Haberman (2008) investigated this method with the value of w fixed at 0.5, but ideally, the value of w should vary with the size of the samples; the larger the samples, the greater the weight for the observed equating. The Bayesian formula of Equation 7.2 offers a solution, for a user who can estimate the sampling variance of the small-sample equating and the variance of the equated scores in the domain. Both of these quantities will vary from one score level to another and not necessarily in proportion with each other. Therefore, with this approach, the value of w in Equation 7.3 would vary from one score level to another.

The variance of the small-sample equating—$\text{var}(y_{current})$ in Equation 7.2—at a given new-form raw-score level—$\text{var}(y_{current})$ in Equation 7.2— is simply the square of the CSEE. There are formulas for estimating this quantity for various equating methods, but the resulting estimates may be highly inaccurate if they are based on data from very small samples of test takers. A possible solution to this problem would be to conduct a series of resampling studies to estimate the CSEE empirically for samples of various sizes.

The variance of the equated scores in the domain of equatings, for a given new-form raw score—$\text{var}(y_{prior})$ in Equation 7.2—can be estimated empirically from prior equatings. The key question is which prior equatings to include in the estimate. We prefer to define the domain of equatings broadly. Limiting the domain to the forms of a single test often narrows the field of prior equatings down to a small sample that may not be representative of a domain that includes the equatings of all future forms of the test. The greatest danger is that the previous forms of a single test may have been much more alike in difficulty than the future forms will be (for an example, see Kim, Livingston, & Lewis, 2008). Limiting the domain of prior equatings to forms of that single test would yield too low a value for $\text{var}(y_{prior})$. The resulting adjustment formula would place too much weight on the identity and too little on the observed equating.

7.6 Introducing the New Form by Stages

Another, very different approach to the small-sample equating problem is to change the way in which new forms of the test are introduced. A new technique, developed by Grant (see Puhan, Moses, Grant, & McHale, 2009) and now being used operationally, is to introduce the new form in stages, rather than all at once. This technique requires that the test be structured in *testlets*, small-scale tests that each represent the full test

Table 7.1 *Plan for Introducing a New Form by Stages*

Form A	Form B	Form C	Form D	Form E	Form F	Form G
Testlet 1	Testlet 7*	Testlet 7	Testlet 7	Testlet 7	Testlet 7	Testlet 7
Testlet 2	Testlet 2	Testlet 8*	Testlet 8	Testlet 8	Testlet 8	Testlet 8
Testlet 3	Testlet 3	Testlet 3	Testlet 9*	Testlet 9	Testlet 9	Testlet 9
Testlet 4	Testlet 4	Testlet 4	Testlet 4	Testlet 10*	Testlet 10	Testlet 10
Testlet 5	Testlet 5	Testlet 5	Testlet 5	Testlet 5	Testlet 11*	Testlet 11
Testlet 6*	Testlet 6	Testlet 6	Testlet 6	Testlet 6	Testlet 6	Testlet 12*

*Not included in computing the test takers' scores.

in content and format. It also requires that the test form given at each administration include one testlet that is not included in computing the test takers' scores.

As an example, consider a test consisting of five testlets that are included in computing the test takers' scores and one additional testlet that is not included in the scores. Table 7.1 shows a plan that might be followed for assembling the first seven forms of this test. The asterisks indicate the testlets that are not included in computing the test takers' scores—one such testlet in each form. With each new form, one of the scored testlets in the previous form is replaced. It is replaced in the printed new form by a new testlet, which is not scored. It is replaced in the scoring of the new form by the testlet that was new (and therefore was not scored) in the previous form.

Each new test form is equated to the previous form in the group of test takers who took the previous form. Form B is equated to Form A in the group of test takers who took Form A, Form C is equated to Form B in the group of test takers who took Form B, and so on. This single-group equating is extremely powerful, because any difference in ability between the sample of test takers and the population is the same for the new form as for the reference form; the equating sample for both forms consists of the same individuals. In addition, the forms to be equated are highly similar, since they have four fifths of their items in common. This overlap of forms limits the extent to which the equating in the sample can deviate from the equating in the population. Because of these two features, this data collection plan is called the single-group, nearly equivalent test (SiGNET) design.

An additional advantage of the SiGNET design is that each new form is equated on the basis of data collected in the administration of the previous form. Therefore, each new form can be equated before it is administered. If a form is administered two or more times before the next form is introduced, the data from those administrations can be combined to provide a larger sample for equating the next form.

Notice in Table 7.1 that Form F is the first form that does not include any of the scored items in Form A. However, Form E has only one fifth of its items in common with Form A, about the same as would be expected if Form E were to be equated to Form A through common items. How does the accuracy of equating Form E to Form A through the chain of single-group equatings in the SiGNET design compare with the accuracy of equating Form E to Form A in a single common-item equating? A resampling study by Puhan, Moses, Grant, and McHale (2009) provides an answer. That study compared the RMSD of equating through a chain of four single-group equatings in a SiGNET design with the accuracy of a single

common-item equating. When all the equating samples in both designs included 50 test takers, the RMSD of the SiGNET equating was about two thirds that of the conventional common-item equating.

7.7 Combining the Approaches

It is certainly possible to combine two or more of the new approaches described above. The circle-arc method, the procedures for using collateral information, and the SiGNET design address different aspects of the equating process. The SiGNET design answers the question, "How should I collect the data for equating?" The circle-arc method answers the question, "How should I use those data to estimate the equating function?" The procedures for incorporating collateral information answer the question, "How should I adjust the estimate to decrease its reliance on the data when the samples are small?"

The possibility of combining these approaches multiplies the number of options available for equating scores on test forms taken by small numbers of test takers. The larger number of possibilities complicates the task of evaluating these procedures. It is useful to know how effective each procedure is when used alone, but it is also useful to know how effective the various combinations of procedures are. To what extent do they supplement each other? To what extent are they redundant? Does the SiGNET design make the use of collateral information unnecessary, or even counterproductive? Would the SiGNET design be even more effective if the single-group equatings were done by the circle-arc method? And, of course, the answers to these questions are likely to depend heavily on the sample size. There are enough research questions here to keep several psychometricians and graduate students busy for a while.

Chapter 7 Appendix

7.A.1 Formulas for Circle-Arc Equating

In the symmetric circle-arc method, the estimated equating curve is an arc of a circle. Let (x_1, y_1) represent the lower end point of the equating curve, let (x_2, y_2) represent the empirically determined middle point, and let (x_3, y_3) represent the upper end point. Let r represent the radius of the circle, and label the coordinates of its center (x_c, y_c).

The equation of the circle is $(X - x_c)^2 + (Y - y_c)^2 = r^2$ or, equivalently, $|Y - y_c| = \sqrt{r^2 - (X - x_c)^2}$. If the new form is harder than the reference form, the middle point will lie above the line connecting the lower and upper points, so that the center of the circle will be below the arc. For all points (X, Y) on the arc, $Y > y_c$, so that $|Y - y_c| = Y - y_c$, and the formula for the arc will be

$$Y = y_c + \sqrt{r^2 - (X - x_c)^2}. \tag{7.A.1}$$

If the new form is easier than the reference form, the middle point will lie below the line connecting the lower and upper end points, so that the center of the circle will be above the arc. For all points (X, Y) on the arc, $Y<y_c$, so that $|Y - y_c| = y_c - Y$, and the formula for the arc will be

$$Y = y_c - \sqrt{r^2 - (X - x_c)^2}. \tag{7.A.2}$$

A simple decision rule is to use Equation 7.A.1 if $y_2>y_c$ and Equation 7.A.2 if $y_2<y_c$.

The formulas for x_c and y_c in the symmetric circle-arc method are a bit cumbersome:

$$x_c = \frac{(x_1^2 + y_1^2)(y_3 - y_2) + (x_2^2 + y_2^2)(y_1 - y_3) + (x_3^2 + y_3^2)(y_2 - y_1)}{2[x_1(y_3 - y_2) + x_2(y_1 - y_3) + x_3(y_2 - y_1)]} \tag{7.A.3}$$

and

$$y_c = \frac{(x_1^2 + y_1^2)(x_3 - x_2) + (x_2^2 + y_2^2)(x_1 - x_3) + (x_3^2 + y_3^2)(x_2 - x_1)}{2[y_1(x_3 - x_2) + y_2(x_1 - x_3) + y_3(x_2 - x_1)]}, \tag{7.A.4}$$

but the formula for r^2 is simply

$$r^2 = (x_1 - x_c)^2 + (y_1 - y_c)^2. \tag{7.A.5}$$

In the simplified circle-arc method, the transformed points to be connected by a circle arc are $(x_1, 0)$, $(x_2, y_2{}^*)$, and $(x_3, 0)$, where

$$y_2{}^* = y_2 - \left(\frac{y_3 - y_1}{x_3 - x_1}\right)(x_2 - x_1). \tag{7.A.6}$$

The transformation of the data points results in a much simpler set of formulas for the coordinates of the center of the circle:

$$x_c = \frac{x_1 + x_3}{2}, \tag{7.A.7}$$

$$y_c = \frac{(x_1^2)(x_3 - x_2) - (x_2^2 + (y_2{}^*)^2)(x_3 - x_1) + (x_3^2)(x_2 - x_1)}{2[y_2{}^*(x_1 - x_3)]}, \tag{7.A.8}$$

and a slightly simpler formula for r^2:

$$r^2 = (x_1 - x_c)^2 + y_c{}^2. \tag{7.A.9}$$

Author Note: Any opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.