Research Report

An Alternative to Equating With Small Samples in the Non-Equivalent Groups Anchor Test Design

Sooyeon Kim

Alina A. von Davier

Shelby Haberman

Research & Development

September 2006
RR-06-27

An Alternative to Equating With Small Samples in the Non-Equivalent Groups Anchor Test Design

Sooyeon Kim, Alina A. von Davier, and Shelby Haberman
ETS, Princeton, NJ

September 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study addresses the sample error and linking bias that occur with small and unrepresentative samples in a non-equivalent groups anchor test (NEAT) design. We propose a linking method called the *synthetic function*, which is a weighted average of the identity function (the trivial equating function for forms that are known to be completely parallel) and a traditional equating function (in this case, the chained linear equating function) used in the normal case in which forms are not completely parallel. Specifically, we compare the synthetic, identity, and chained linear functions for various-sized samples from two types of national assessments. One design uses a high reliability test and an external anchor, and the other uses a relatively low reliability test and an internal anchor. The chained linear equating functions derived from the total sample are used as the criterion equating function in both assessments. The results from each of these methods were compared to the criterion equating function with respect to linking bias and error. The study indicates that the synthetic functions might be a better choice than the chained linear equating method when samples are neither large nor representative.

Key words: Linking bias, sample error, equating functions, NEAT design, synthetic function, identity

Acknowledgments

This paper was originally prepared for *Recent Advances in Score Equating,* a symposium at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, April 7–11, 2006. We would like to thank Neil Dorans, Anna Kubiak, Kevin Larkin, Jonathan Manalo, Alan Shaw, and Michael E. Walker for their many helpful comments on earlier drafts of this paper.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Test equating is a statistical method that makes scores from different test forms interchangeable by adjusting for differences in difficulty among forms built to the same specifications. As with other statistical procedures, the equating of test scores is subject to sampling effects such as sampling error or sampling bias. If the sample is large and representative, the equating relationship in the sample may accurately represent the equating relationship in the population. The smaller the sample, the more likely it is that the equating function computed for that particular sample differs substantially from that of the population. Additionally, sampling error may affect the extent to which the sample represents the population from which it was drawn, and, as a result, can influence the quality of the equating. In practice, the use of a small sample can be expected to have more influence on equating when equated samples are not representative.

We propose a linking method called the *synthetic linking function*, which can be used instead of conventional test equating when samples are very small and unrepresentative. As the name indicates, the synthetic function is a weighted average of the identity function appropriate for equating when forms are completely parallel and a traditional equating function (e.g., the chained linear equating function). This function has not been defined in the literature. The major purpose of this study is to investigate the potential benefits of the synthetic method for linking that involves small samples. This study addresses the sample error and linking bias that occur with small samples in the non-equivalent groups anchor test (NEAT) design with respect to various equating/linking functions. Identity, chained linear, and synthetic functions are compared to the criterion function in terms of linking error resulting from sample variability and bias for different sized samples. The same form of analysis is conducted using two types of national assessment programs, one with a highly reliable external anchor and one with a moderately reliable internal anchor, to enhance the generalizability of the findings.

The analysis is preliminary to the extent that, in the synthetic function, the weighting of components is even rather than selected on any rational grounds. Thus the present study illustrates the potential benefit of a synthetic function, but it does not provide an operational procedure.

Small-Sample Equating

Most of the literature on small-sample equating focuses on the minimum sample size needed to ensure the accuracy of test equating. For example, Hanson, Zeng, and Colton (1991) examined the functioning of various equating methods with samples ranging from 100 to 3,000 observations. They found that, with samples of 100, the identity function provided lower linking errors (including bias and sample error, which is zero for the identity function) than any other linear and equipercentile equating method in the random groups equating design. According to Harris (1993) and Kolen and Brennan (2004), a rule of thumb is a sample size of at least 400 per form when using linear equating or equating based on the Rasch item response theory model, and 1,500 per form when using equipercentile equating or equating based on a three-parameter item response theory model. In any given situation, however, the shapes of the distributions and the degree of linking precision required should be examined when considering sample size requirements. In addition, in the case of equipercentile equating, the effects of log-linear smoothing (Holland & Thayer, 2000) require consideration. Because linking may be situation specific, the literature provides no definitive recommendation for sample sizes for an appropriate linking process.

In related research, Livingston (1993) examined the effectiveness of log-linear models used to presmooth discrete probability distributions. He considered samples of 10, 25, 50, 100, and 200 examinees per form in the common-item NEAT design. He found that the benefits of smoothing were greatest when the sample was small, but that the number of moments in the observed distribution that should be preserved in the smoothed distribution depends on the sample characteristics. It is also worth noting that if the samples are small, then log-linear smoothing, a nonlinear process, may introduce sampling bias that can offset the gain due to reduction in the standard error of equating (SEE).

Based on a slightly different perspective, von Davier and Kong (2005) examined the SEE and the statistical sensitivity of the standard error of equating difference (SEED) in the NEAT design among linear equating methods for different size samples. They concluded that differences between the Tucker and chained linear equating functions might not be statistically significant with a sample of 100 observations, although they may be significant in the original large samples. This finding indicated that substantial uncertainty results when computing equating functions with a small sample. Additionally, they found that a sample of 200

observations was apparently large enough for the SEED to detect a statistical difference between the two methods for most of the score range. The SEEs and the SEEDs are asymptotic results; it is therefore important to determine how they vary with sample size. For samples as large as 10,000 observations, parameters will be estimated accurately; consequently, the ±2SEE and ±2SEED bands will be narrow. As samples become smaller, these bands become broader, indicating greater uncertainty in the equating results, especially at the extreme scores.

Parshall, Houghton, and Kromrey (1995) examined the effects of sample size on the stability and bias of linear equating with two parallel forms based on the common-item non-equivalent-group design. Specifically, they examined the SEE and statistical bias in linear equatings with small samples, such as 15, 20, 50, and 100. Their results suggested trivial levels of equating bias even with small samples but substantial increases in standard errors as sample size decreases. The sampling error is smallest in the proximity of the mean raw score of the examination, and the error increases monotonically (but not linearly) as a function of the deviation of scores from the mean. This implies that the standard error associated with differences in sample size becomes more pronounced for scores at greatest distances from the mean raw score for the examination.

In a similar manner, Skaggs (2005) used an equivalent-groups design to study the equating of the passing score on a certification test using samples ranging from 25 to 200 observations. He selected four equating methods—linear, mean, presmoothed (using log-linear models), equipercentile, and unsmoothed equipercentile—and compared them with respect to equating bias and the SEE. He found that the SEE became smaller as sample size increased; however, equating bias changed little as a function of sample size. Even the sample that included 200 observations evinced substantial equating error on at least part of the raw score scale, yielding a significant percentage of misspecified examinees in terms of their pass/fail designations. Additionally, Skaggs found that two- and three-degree smoothing generally produced smaller standard errors than did unsmoothed equating, and, as Livingston (1993) found, this benefit increased as sample size decreased to, but not below, 50 observations.

As summarized above, several studies have been conducted to examine the effects of small samples on equating with respect to equating error or statistical bias. Some experts believe that use of an identity function is preferable to nontrivial equating with extremely small samples, as the large random equating error associated with very small samples negates the benefits of

equating. Statistical methodology cannot solve problems caused by poorly collected data or data insufficiency. In many testing programs (e.g., in programs with a consistently low volume of examinees), however, it may not be feasible to obtain large samples. Nevertheless, these programs need to report, in a timely manner, comparable scores over different administrations or test forms. This dilemma is particularly severe in the common situation in which form construction has not been well-structured.

Identity Function Versus the Weighted Function

As mentioned, an alternative to equating with small samples is to use the *identity function*. In general, the identify function will be appropriate when test specifications are well-defined and two forms are nearly parallel in both difficulty and content. Harris (1993) pointed out two reasons for not conducting equating: (a) equating is unnecessary or (b) equating is unwarranted. In practice, equating is unnecessary when the forms to be equated are similar enough that equating would likely produce more error than it would remove. Additionally, Harris stated that no equating should be considered when samples are small. Recently, Skaggs (2005) found that when using samples as small as 25 no equating is likely to do less harm to examinees than some form of equating because equating with small samples may lead to a degree of equating error that could exceed the total equating error variance, at least when using linear equating methods.

Whether or not to use the identity function will be a major decision in a small sample situation. This decision may depend on several variables such as specific sample size, location of passing scores on the raw score scale (if applicable), equating design (e.g., NEAT or random group design), and degree of difference between forms. For example, Skaggs (2005) recommended the use of identity function in the random group case when forms differ by one-tenth of a standard deviation or less, indicating a small difference between them. Using the identity function may allow a small amount of equating error when test forms are carefully developed from the same set of test specifications. Ironically, well-controlled test assembly for a stable test is usually accompanied by ample data available for equating. Data paucity is likely to affect the test assembly process as well as the equating process (Neil Dorans, personal communication, May 8, 2006). This means that the use of the identity function instead of some form of equating may lead to a large amount of unknown bias, called systematic error.

In general, the total equating error can be partitioned into random error and systematic error components. Which is more worrisome in small sample equating situations: bias or equating error (i.e., random error) resulting from sample variability? The answer may depend on the situation. For example, bias may be problematic especially for tests with cut scores. The sample equating function may have less linking bias than the identity function; however, for small samples, it has quite a bit of error due to sample variability. Conversely, the identity is usually quite biased but it has no sample variability (i.e., error). The intent in using the identity is for the increase in systematic error to be more than offset by the decrease in random error. The sum of random equating error variance and squared bias equals the mean squared error in equating. Based on this relationship, if the squared bias associated with the identity function is less than the random equating error variance associated with using other equating methods, the identity function is preferable to other equating functions (Kolen & Brennan, 2004, p. 289).

This study is designed to introduce an alternative, called the synthetic linking function, which is a weighted average of the identity function (having no equating error but large bias) and the traditional equating function (having small bias but large equating error). The synthetic function is not normally an equating function, for it does not in general satisfy the invertibility requirement for equating functions. However, this requirement may be achieved with an appropriate weight system. Unless the weight attached to the identity function decreases to 0 as the sample size increases, the synthetic function need not converge to the population equating function as the sample size increases.

Appropriate weighting of the identity and conventional equating functions is not clear without either historical information concerning variability of form difficulty or specific information concerning construction of forms. Despite these limitations related to the synthetic function, the synthetic function can be a better choice than the sample equating function or the identity function in certain specific contexts. For instance, with very small equating samples, the synthetic function may provide smaller linking error than other traditional methods, and it may also provide less bias than the identity function does. The synthetic function offers a useful compromise between the two and can better reduce the total mean squared error than can the sample equating function or the identity function when equating samples are small. For that reason, we suggest the synthetic function as an alternative where test equating must be based on

very small samples; however, the proper development of weights is a subject for further research. The equations for the synthetic function are presented in detail in the methodology section.

The Major Purpose of This Study

This study addresses the sample error and equating bias that occur with small and unrepresentative samples in the NEAT design. Specifically, this paper focuses on identity, chained linear, and synthetic functions with samples of various sizes (N = 10, 25, 50, 100, and 200) from two types of national assessments: one with a highly reliable external anchor (Study 1) and one with a moderately reliable internal anchor (Study 2). Study 2 uses data sets from a teacher licensure testing program that requires pass/fail designations. Because the original samples from these assessments are large, the chained linear equating function derived from the total samples is used as the criterion for both types of assessment. We compare the results from the three methods (identity, sample-based chained linear, and synthetic functions) with the criterion function with respect to equating bias, equating error, and the root mean squared error (RMSE) index, defined as a function of equating error and bias.

Methodology

Design

A NEAT design is often used with small samples in practice. As presented in Table 1, in the NEAT design there are two test forms, X and Y, to be equated and a target population, T, for which the equating is done. These two operational tests (X and Y) are given to two samples of examinees from different test populations or administrations (denoted here by the populations P and Q). Accordingly, the two test scores, X and Y, are each only observed *either on P or on Q, but not on both*. An anchor test, V, is given to both samples from P and Q resulting in the following data structure, where ✓ denotes the presence of data. The anchor test score, V, can be either a part of both X and Y (the internal anchor case) or a separate score (the external anchor case).

In the NEAT design, the Target Population, T, is a *mixture* of P and Q in which P and Q are regarded as partitioning T. P and Q, are also given weights that sum to 1. This is denoted by

$$T = \kappa P + (1-\kappa)Q\,. \tag{1}$$

The partition of T is determined by the weight κ. If $\kappa = 1$, then $T = P$ and if $\kappa = 0$ then $T = Q$. If $\kappa = ½$, then P and Q are represented equally in T. Any choice of κ between 0 and 1 is possible and reflects the amount of weight that is given to P and Q in determining the target population.

Table 1

Data Collection: NEAT Design

Target population	Original populations	X	V	Y
T	P	√	√	
	Q		√	√

Equating/Linking Functions Used in This Study

As mentioned earlier, we investigated only the chained linear equating method (equating Form X to Form Y) in this study, along with the identity and synthetic functions, because we expected similar equating results from other linear methods (e.g., Tucker, Levine), and the samples of interest are too small to ensure the adequacy of equipercentile equating results (Harris, 1993; Kolen & Brennan, 2004). The description for these methods is as follows:

Chained linear equating function. Many observed score equating methods are based on the linear equating function. Usually, the rationale behind linear equating on the target population, T, is to set standardized deviation scores (z-scores) on the two forms to be equal such that

$$\frac{x-\mu_{XT}}{\sigma_{XT}} = \frac{y-\mu_{YT}}{\sigma_{YT}},$$

where $\mu_{YT}, \sigma_{YT}, \mu_{XT},$ and σ_{XT} are the means and the variances of X and Y in T.[1] Solving for y in the above equation results in the equation for the linear equating function:

$$Lin_{YT}(x) = \mu_{YT} + \sigma_{YT}\left[\frac{(x-\mu_{XT})}{\sigma_{XT}}\right]. \quad (2)$$

In this study, we use the chained linear equating function as a generic linear function in the NEAT design. We expect similar results for the other linear observed-score equating functions in the NEAT design, such as the Tucker and the Levine observed-score equating functions. All these functions and their (untestable) assumptions are described in detail elsewhere (von Davier & Kong, 2005; Kolen & Brennan, 2004).

The chained linear equating function in the NEAT design is given by chaining the two linear linking functions (i.e., by using the mathematical composition of the two linear functions), from X to V on P [$\mathrm{Lin}_{VP}(x)$] and from V to Y on Q [$\mathrm{Lin}_{YQ}(v)$]. This results in the usual form for the chain linear equating function as follows:

$$\begin{aligned} CL_{YT}(x) &= \mathrm{Lin}_{YQ}\left(\mathrm{Lin}_{VP}(x)\right) \\ &= \mu_{YQ} + \left(\sigma_{YQ}/\sigma_{VQ}\right)\left(\left(\mu_{VP} + \left(\sigma_{VP}/\sigma_{XP}\right)\left(x-\mu_{XP}\right)\right) - \mu_{VQ}\right) \\ &= \mu_{YQ} + \left(\sigma_{YQ}/\sigma_{VQ}\right)\left(\mu_{VP}-\mu_{VQ}\right) + \left(\sigma_{YQ}/\sigma_{VQ}\right)\left(\sigma_{VP}/\sigma_{XP}\right)\left(x-\mu_{XP}\right), \end{aligned} \tag{3}$$

von Davier, Holland, and Thayer (2004) showed that the chain linear function from Equation 3 can be written in the same general form as Equation 2.

Identity linking function. An alternative to equating with small samples is to not conduct any equating at all. Formally, the identity function is

$$ID_{YT}(x) = x, \tag{4}$$

where x is a raw score of the new form X that is placed on the raw-scale of the old form Y. Obviously, the identity function is linear. The random equating error is zero for the identity linking, because the equated scores are obtained through a deterministic procedure. As mentioned before, however, the use of the identity can increase systematic error (i.e., bias). For the identity linking to function properly, the two test forms to be equated need to have the same number of items (i.e., the same number of possible scores).[2]

Synthetic linking function. We propose an alternative method called the synthetic linking function for equating with small samples. The synthetic linking function creates a compromise between the identity function (no equating) and linear linking by combining them using a specific weight system:

$$Syn_{YT}(x) = w \times CL_{YT}(x) + (1-w) \times ID_{YT}(x), \quad (5)$$

where w is a weight between 0 and 1; x is raw score in Form X; CL_{YT} is the chained linear function (which can be replaced by other types of equating functions); and ID_{YT} is the identity function. Note that the weight w from Equation 5 is not the same as the weight κ from Equation 1 that is used for defining the target population.

As discussed previously, equating with small samples may lead to large sampling error reflected in the SEE. Given that the SEEs for the identity linking are zero, the SEEs for the synthetic linking can be substantially reduced as compared to the chained linear function used here:

$$\begin{aligned} \text{Var}(\text{Syn}_{YT}(x)) &= w^2\text{Var}(CL_{YT}(x)) + (1-w)^2\text{Var}(ID_{YT}(x)) + 2w(1-w)\text{Cov}(CL_{YT}(x), ID_{YT}(x)) \\ &= w^2\text{Var}(CL_{YT}(x)). \end{aligned} \quad (6)$$

Hence,

$$SEE(Syn_{YT}(x)) = .5 \times SEE(CL_{YT}(x)) \quad (7)$$

if $w = .5$.

As can be seen, when giving the same weight to the identity and chained linear functions, the SEE is reduced by one half. Similarly, Equation 8 shows that the bias of the synthetic function that is mostly introduced by use of the identity function can be reduced under the assumption that the chained linear equating function is not biased or much less biased when compared to the identity function.

$$\mu(syn_{YT}(x)) = w \times \left[\mu(CL_{YT}(x))\right] + (1-w) \times \left[\mu(ID_{YT}(x))\right]. \quad (8)$$

One issue related to use of the synthetic function has to do with the manner in which the two functions are averaged using a certain weight system. There is no literature/definitive guidance regarding the weight systems when averaging the two different equating/linking functions. The best approach in this case will be to reevaluate the data-collection design, beginning with the design of the test itself (Michael Walker, personal communication, May 3, 2006). However, more often the weight system for the two functions will be investigated after

collecting data. In that case the following aspects can be considered as some guidelines to deciding on specific weights: the a priori information about the tests forms and the anchor (e.g., sample size, tests and anchor reliability, test specifications, and test variability over time) and information about the two populations of examinees. For example, when the standardized mean differences on the anchor in the two equating samples are small (< 0.1) but the mean difficulty difference of the test forms is large (> ½ standard deviation), the identity function should have a low weight (less than .5 in Equation 5). In other words, the identity function should be weighted less because it cannot adjust for the difference in difficulty between the two forms, which would introduce bias in the equating results and because the conventional equating function can be determined with relatively good accuracy. Conversely, when the mean difficulty difference of the two tests forms is small, the identity function can be highly weighted (greater than one half in Equation 5). That is, the identity function might receive a higher weight (closer to 1) when test forms are constructed to be nearly parallel. The weight on the identity function might decrease when the sample size increases.

Although weight systems for the two functions are flexible, a simple scheme to illustrate use of the synthetic function weights the two functions equally. This weighting will be used in this study. Based on the finding about test variability, however, other weights might be used as well. A more thorough analysis for choosing the appropriate weights is an interesting issue for further research, and it will be presented elsewhere. An equation that will transform the ordinary weight system into the symmetric weight system is presented in the appendix.

Criterion function. We chose the linear equating derived from the substantially large total samples as the equating criterion. It may be questionable why the total sample criterion function is linear in nature, but it is worth noting that the linear function is the linear part or first term in an expansion of the equipercentile function (von Davier, Holland, & Thayer, 2004, p. 12). The remainder (called $R(x)$) caused by the shape difference between linear and nonlinear functions is small if the score distributions of the tests are similar in shape. We chose the linear criterion to avoid confounding the differences in accuracy with the differences in the shape between the large and small samples' equatings. We compared the chained linear, identity, and synthetic functions in small samples with a chained linear equating function derived from the total groups with respect to equating bias, SEE, and an RMSE index.

Procedure

As the first step, we used the chained linear method with the total samples from *P* and *Q* and then obtained the Form *X* scores equated to the Form *Y* scores. These equated *X* scores are the equating criterion. As the next step, we linked the Form *X* scores to the Form *Y* for the five sample sizes. For each sample size, we estimated bias and RMSE by use of a Monte Carlo simulation with 1,000 replications. For each replication and sample size, the three different equating/linking methods were then applied. These small sample equating results were compared with the equating criterion derived from the total groups by computing differences over the entire range of scores. The differences between them were investigated with respect to the following deviance measures.

Deviance Measures

In order to evaluate equating/linking results, equating bias (see Equation 9) and the SEE (see Equation 10) were calculated over 1,000 replications, along with the RMSE and standardized root mean squared error (SRMSE) that is free from the score scale (see Equations 11 and 12). Particularly, the SRMSE is used for comparing the results from assessments having different reporting scale lengths.

Equating bias was defined as the mean difference between an equating method and criterion equating over 1,000 replications. The standard deviation of these differences over 1,000 replications is a measure of the SEE or error due to sample variability. The sum of squared bias and squared SEE is an indication of total equating error variance, and the square root of this value is an RSME index. The following equations represent bias, equating error, and RMSE measures conditioned on each raw score point (x_i):

$$Bias_i = \bar{d}_i = \frac{\sum_{j=1}^{J}\left[\hat{e}_{yj}\left(x_i\right) - e_y\left(x_i\right)\right]}{J}. \tag{9}$$

$$SEE_i = sd_i = \sqrt{Var_j\left[\hat{e}_{yj}\left(x_i\right) - e_y\left(x_i\right)\right]} = \sqrt{Var_j\left[\hat{e}_{yj}\left(x_i\right)\right]}. \tag{10}$$

$$RMSE_i = \sqrt{\bar{d}_i^2 + sd_i^2}. \tag{11}$$

$$SRMSE_i = \frac{\sqrt{\overline{d}^2 + sd^2}}{\sigma_{YQ}}, \quad (12)$$

where i is the number of score points, j is the number of replications (J = 1,000), $\hat{e}_{yj}(x)$ denotes the raw score equivalent calculated from one of the equating functions (chained linear, identity, and synthetic functions) in the sample j, $e_y(x)$ indicates the raw score equivalent from the criterion equating function, and σ_{YQ} is the standard deviations of Y in Population Q.

As shown in these equations, bias indicates the difference between the average estimated y-score equivalent across all samples and the population (actual) y-score equivalent, and equating error indicates the difference between the estimated y-score equivalent for a sample and the average estimated y-score equivalent across all samples. These conditional bias, equating error, and RMSE indexes were averaged over the entire range of raw score points, respectively, to obtain a single summary measure for each deviance measure.

Small-Sample Selection

New samples of 10, 25, 50, 100, and 200 observations were randomly selected from P (those who took (X, V)) and Q (those who took (Y, V)), respectively, using SAS PROC SURVEYSELECT. This method of simple random sampling selects units without replacement. Each possible sample of n units drawn from N has the same probability of being selected. For each sample size, we randomly selected 1,000 samples with the same sampling rate and sample design. Accordingly, 5,000 (= 5 × 1,000, where five represents the five sample sizes) random samples were drawn from the original N_P for (X, V) and N_Q for (Y, V); samples were then paired to equate Form X to Form Y using the linking/equating methods**.** The exact same sampling procedures were employed for the real data sets from two types of national assessments.

The next two sections describe two studies with similar formats conducted with the two data sets of the different testing programs. The first study uses an *external* anchor in the NEAT design; the reliabilities for both the tests (.92 – .94) and the anchors (.84) are high. Although anchors are external in this design, the correlations between the tests and the anchors (.87 – .88) are still high. This suggests more weight be placed on the identity function (see Table 2 for details). The second study uses an *internal* anchor in the NEAT design, and the reliabilities for

both tests (.82 – .86) and anchors (.67) are lower than those in the previous study. Although internal anchors were used, the correlations between the tests and the internal anchors (.88) are about the same as in the external anchor design (see Table 5). The data sets used in Study 2 are from a licensure testing program having cut scores. Those cut scores (which fall in the exact raw score range from 64 to 77) are located near the mean, suggesting less weight be placed on the identity function. Although we expect the identity function to perform differently depending on data characteristics, the equal weight system (.5) was employed when synthesizing the two functions in both studies. Consequently, these studies may tell us how the identity function works differently as a function of test reliability and correlations between total test and anchor in practice.

Study 1

Data

This study used data from two national administrations of a high-volume testing program. The administrations took place in the fall of 2001 (called *P*) with Form *X* and in the winter of 2000 (called *Q*) with Form *Y* (the same set of data as in von Davier & Kong, 2005). The data were collected following a NEAT design with an external anchor, as shown in Table 1. The data sets consist of the raw sample frequencies of rounded equation scores[3] for two parallel, 78-item tests (*X* and *Y*; $x_1 = y_1 = 0,\ x_2 = y_2 = 1,\ \cdots,\ x_{79} = y_{79} = 78$) and a 35-item external anchor test (*V*; $v_1 = 0,\ v_2 = 1,\ \cdots,\ v_{36} = 35$) given to two samples (*P* and *Q*) from a national target population of examinees (*T*). The number of examinees was 10,634 in *P* and 11,321 in *Q*. Descriptive statistics for these groups are summarized in Table 2.

Table 2

Summary Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 1

	N	μ	σ	SEM	Reliability	ρ
X_P	10,634	39.25	17.23	4.5 – 4.6	.91 –.94	.88
V_P		17.05	8.33	3.3	.84	
Y_Q	11,321	32.69	16.73	4.5 – 4.6	.91 –.94	.87
V_Q		14.39	8.21	3.3	.84	

Note. SEM = Standard error of measurement. ρ = Correlation between total score and anchor.

As shown in Table 2, the mean of the anchor test *V* is 17.05 (±0.08) in population *P* and 14.39 (±0.08) in *Q*, where 0.08 is the standard error of the mean. Thus, population *Q* is less proficient than population *P*, as measured by *V*. In terms of effect sizes, the difference between these two means (2.66) is approximately 32% of the average standard deviation of 8.27. This magnitude indicates a fairly large difference between the two populations for this type of testing program. However, psychometric properties (e.g., standard error of measurement, reliability, correlation between total score and anchor) for the two forms were almost identical across administration groups.

Small samples of 10, 25, 50, 100, and 200 observations were randomly selected from *P* (those who took (*X*, *V*)) and *Q* (those who took (*Y*, *V*)) and paired to be used as equating samples according to the procedures described above. The summary statistics for *X, Y,* and *V* in *P* and *Q* over 1,000 pairs of samples are presented in Table 3, along with statistics for the total group.

Results

Form *X* was equated to Form *Y* using a chained linear equating method with a sample of 10,634 examinees for Form *X* and 11,321 examinees for Form *Y*. For each raw score (0 to 78) on Form *X*, the equivalent raw score on Form *Y* was determined. This raw-to-raw conversion served as the equating criterion. To examine the effectiveness of small-sample equating, three linking/equating methods were examined: chained linear, the identity function, and the synthetic link. The results from each of these methods were compared to the equating criterion by subtracting the criterion raw score equivalents from the raw score equivalent produced by one of the equating methods for each possible raw score on Form *X*. The differences between them were summarized with respect to equating bias, SEE, and RMSE/SRMSE indexes. Table 4 presents the summary statistics averaged over the entire range of raw scores within each combination of sample size and equating approach.

As shown in Table 4, both equating bias and error became smaller as sample size increased. By design, the averaged difference (i.e., bias) between the identity function (no equating) and the criterion equating on the total population was constant across the five sample sizes. As expected, the identity function showed the greatest amount of equating bias (1.171) compared to the chained linear and synthetic functions, but there is no equating error for this function.

Table 3

Summary Statistics for the NEAT Design With a Highly Reliable External Anchor for Five Different Sample Sizes

	μ_{XP}	σ_{XP}	μ_{VP}	σ_{VP}	μ_{YQ}	σ_{YQ}	μ_{VQ}	σ_{VQ}	ρ_{xv}	ρ_{vy}
Total	39.25	17.22	17.05	8.33	32.68	16.72	14.38	8.20	.88	.87
					$N = 10$					
Average	39.48	16.79	17.16	8.18	32.78	16.38	14.38	7.97	.87	.86
SD	5.58	3.36	2.69	1.58	5.22	3.16	2.65	1.54	.11	.11
Max	55.80	27.17	25.40	13.61	48.90	27.02	22.00	12.73	.99	.99
Min	21.70	5.83	7.00	2.11	17.80	5.64	7.00	3.73	-.10	-.02
					$N = 25$					
Average	39.16	17.08	16.97	8.27	32.63	16.66	14.34	8.16	.88	.87
SD	3.44	2.01	1.67	.94	3.28	2.03	1.65	.96	.05	.06
Max	51.72	23.48	22.24	10.99	43.96	23.80	19.68	12.37	.98	.97
Min	27.48	10.70	11.28	5.58	21.08	9.03	9.24	5.22	.57	.46
Total	39.25	17.22	17.05	8.33	32.68	16.72	14.38	8.20	.88	.87
					$N = 50$					
Average	39.21	17.12	17.02	8.29	32.72	16.69	14.39	8.19	.88	.87
SD	2.40	1.39	1.13	.65	2.35	1.35	1.16	.65	.04	.04

(Table Continues)

Table 3 (continued)

	μ_{XP}	σ_{XP}	μ_{VP}	σ_{VP}	μ_{YQ}	σ_{YQ}	μ_{VQ}	σ_{VQ}	ρ_{xv}	ρ_{vy}
Max	48.44	21.25	21.70	10.04	41.04	11.81	18.08	10.27	.95	.95
Min	30.16	12.83	13.90	6.14	24.90	21.68	10.60	6.24	.63	.66
					$N = 100$					
Average	39.17	17.24	17.03	8.33	32.64	16.68	14.36	8.19	.88	.87
SD	1.71	.97	.82	.46	1.64	.94	.79	.44	.03	.03
Max	45.12	20.21	19.86	9.67	38.13	19.28	16.69	9.55	.94	.93
Min	33.65	13.94	14.38	6.74	27.62	13.70	11.80	6.93	.64	.72
					$N = 200$					
Average	39.28	17.20	17.06	8.32	32.62	16.72	14.35	8.20	.88	.87
SD	1.24	.65	.59	.31	1.16	.64	.56	.31	.02	.02
Max	42.57	19.26	18.96	9.27	36.71	19.12	16.45	9.17	.93	.92
Min	35.32	15.23	15.06	7.40	29.59	14.87	12.91	7.25	.79	.80

Note. μ_{XP}, σ_{XP}, μ_{VP}, and σ_{VP} are the means and the standard deviations of X and V in P. μ_{YQ}, σ_{YQ}, μ_{VQ}, and σ_{VQ} are the means and the standard deviations of Y and V in Q. ρ_{xv} is the correlation between X and V in P. ρ_{vy} is the correlation between Y and V in Q. The average, SD, max, and min indicate mean, standard deviation, maximum value, and minimum values for each statistics (e.g., μ_{XP}, σ_{XP}, μ_{VP}, σ_{VP}, and ρ_{xv}), respectively, computed over 1,000 replications.

Table 4

Summary of Differences Among Various Equating Functions Using Small Samples in a NEAT Design With a Highly Reliable External Anchor

Deviance measures	Chained linear–criterion	Identity–criterion	Synthetic–criterion
		$N = 10$	
Bias	.989	1.171	.877
SEE	7.032	.000	3.516
RMSE	7.101	1.171	3.623
SRMSE	.425	.070	.217
		$N = 25$	
Bias	.241	1.171	.601
SEE	3.915	.000	1.957
RMSE	3.922	1.171	2.048
SRMSE	.235	.070	.123
		$N = 50$	
Bias	.114	1.171	.607
SEE	2.609	.000	1.305
RMSE	2.612	1.171	1.439
SRMSE	.156	.070	.086
		$N = 100$	
Bias	.019	1.171	.594
SEE	1.777	.000	.889
RMSE	1.777	1.171	1.069
SRMSE	.106	.070	.064
		$N = 200$	
Bias	.012	1.171	.589
SEE	1.262	.000	.631
RMSE	1.262	1.171	.863
SRMSE	.075	.070	.052

Note. The values for the bias, SEE, RMSE, and SRMSE were averaged across the entire range of raw scores that were calculated over 1,000 replications.

The chained linear method showed the smallest bias (from .012 to .989), but it also showed the greatest amount of equating error (from 1.262 to 7.032). The synthetic function showed the next highest level of bias (from .589 to .877) and showed the next highest equating error (from .631 to 3.516) across the five sample sizes. As Equation 7 indicates, the SEE of the synthetic function was reduced by one-half compared to the SEE of the chained linear function. As a function of equating bias and error, RMSE and SRMSE also showed a consistent pattern across the five sample sizes. RMSE and SRMSE were smaller for all methods as sample size increased. The identity function showed the smallest RMSE compared to other functions for samples as small as 50. When sample size increased, however, the synthetic function showed a smaller RMSE than the identity function. The RMSE index of the chained linear function was much larger than that of other functions, even when samples were as large as 100. In addition, the identity function showed a SRMSE of equating less than .1. The SRMSE for the synthetic function was less than .1 for samples as small as 50, but chained linear approached this level only for samples of 100 or greater. At samples of 200, all the methods resulted in SRMSE values of less than .1.

Figures 1 to 5 plot the bias of equating for each method compared to the criterion at each score point. As shown in these figures, chained linear equating appeared to be better than the other techniques when it was averaged over 1,000 replications. Even for samples of size 25, chained equating showed very little bias for the middle raw score range, 35 to 55, although the bias increased toward both ends of the raw scores. As presented in Figure 5, chained linear showed almost no bias across the entire range of raw scores when samples were as large as 200. The average bias of chained linear equating was located within one SEE criterion even when samples were as small as 50. The identity function tended to produce higher raw equivalent scores compared to the criterion, and this tendency was stronger when raw scores increased. As expected, the bias of the synthetic function was located somewhere between chained linear and identity across the entire range of raw scores.

Because equating bias was averaged over 1,000 replications, it might misrepresent the actual trend. For that reason, the bias of chained linear equating based on small samples was presented along with its error band representing plus or minus one empirical SEE. This range represents the 68% confidence interval for the chained linear equating function. In both cases, the 68% band for chained linear equating was very wide, implying severe fluctuation across 1,000 replications. As expected, the 68% band became narrower as sample size increased. As

presented in Figure 4, the 68% band was still wide even for the medium sample size (e.g., 100) condition, but the chained linear function might be as effective as the identity function for the raw score region 35 to 50. As presented in Figure 5, chained linear equating approached a level of bias and error similar to that of the identity and synthetic functions with samples of 200. The 68% band for the synthetic function was also presented in these figures. As Equation 7 indicates, this band was located in the middle of the 68% band for the chained linear function.

The comparison between the chained linear function and the synthetic function was particularly interesting. When sample sizes were small (less than 50), in general, the synthetic function outperformed the chained linear function because the identity function (no equating error) is part of the synthetic function. The figures suggest that, for samples as small as 10, 25, or 50, the identity function is generally preferable to any of the other linking functions under study. This result corroborated Skaggs' (2005) finding. Although Skaggs concluded that some form of equating is clearly preferable to no equating for samples larger than 75, the identity function (no equating) showed a generally smaller RMSE index than the chained linear function for samples as large as 200 in the current study. The reason may be related to test design in that Forms *X* and *Y* were close to parallel. However, the synthetic function, which contains a chained linear component, showed a smaller RMSE than did the identity for samples as large as 100.

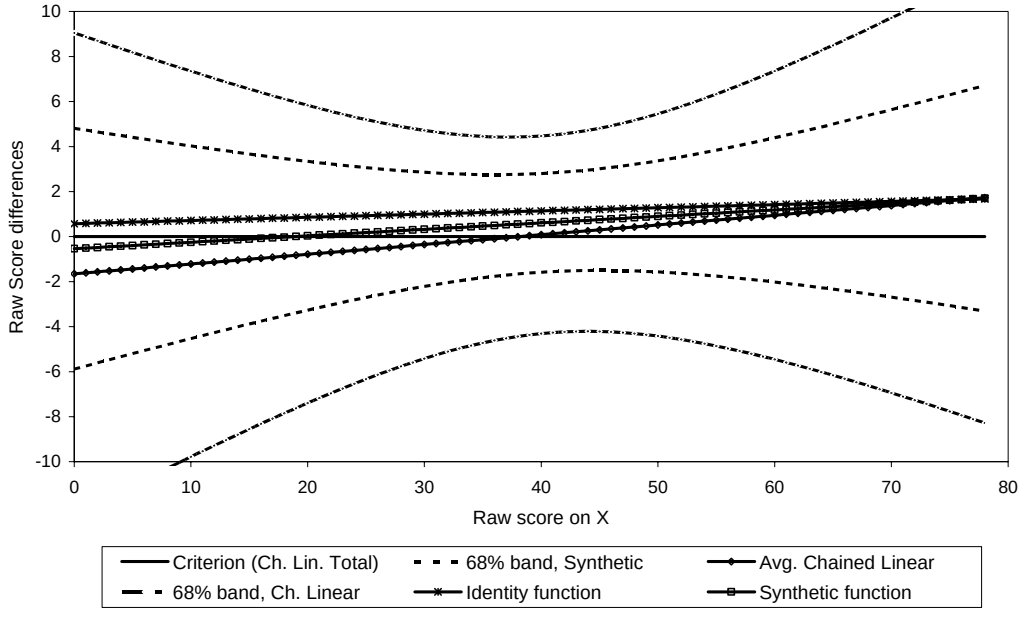

***Figure 1*. Equating bias of different methods over 1,000 replications (*N* = 10) in the high-reliability external anchor design.**

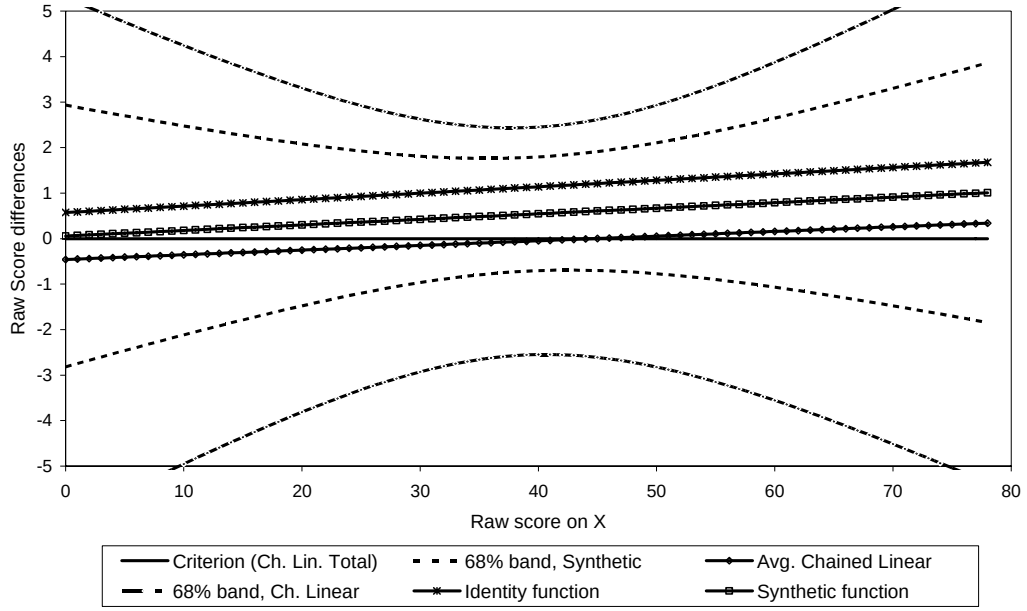

Figure 2. **Equating bias of different methods over 1,000 replications (*N* = 25) in the high-reliability external anchor design.**

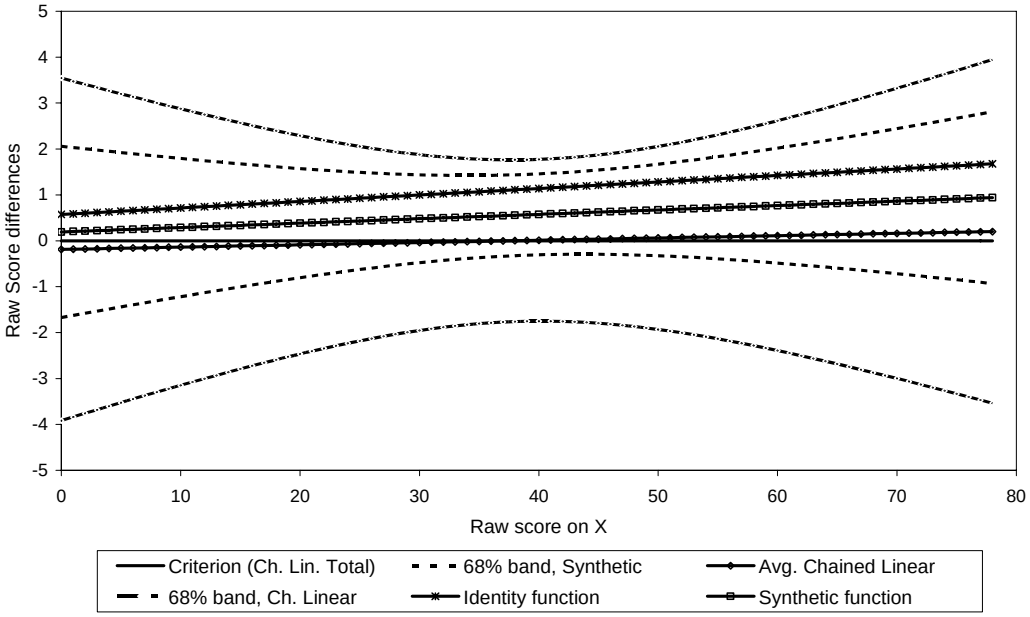

Figure 3. **Equating bias of different methods over 1,000 replications (*N* = 50) in the high-reliability external anchor design.**

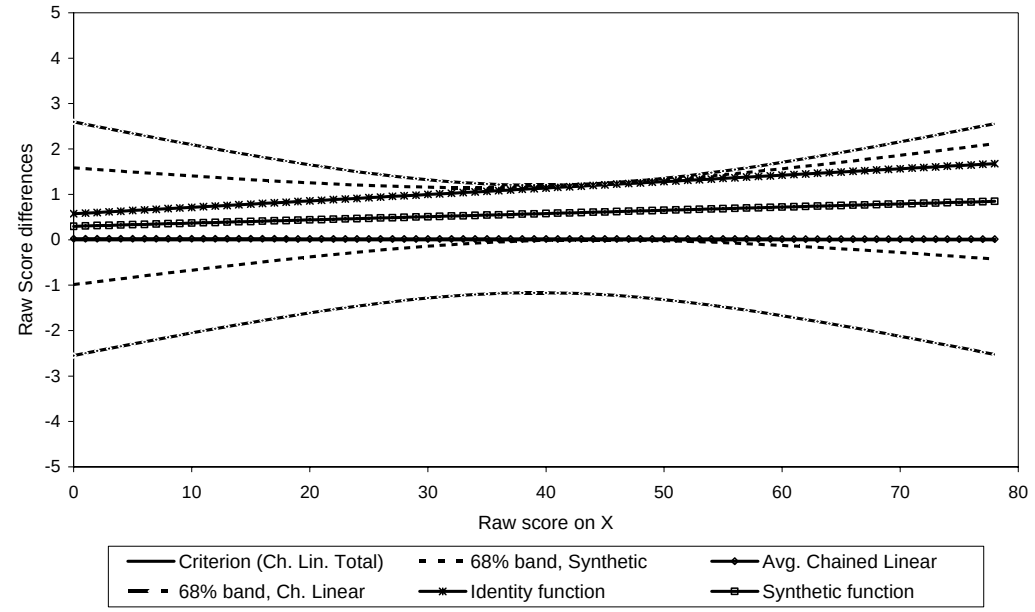

Figure 4. **Equating bias of different methods over 1,000 replications (*N* = 100) in the high-reliability external anchor design.**

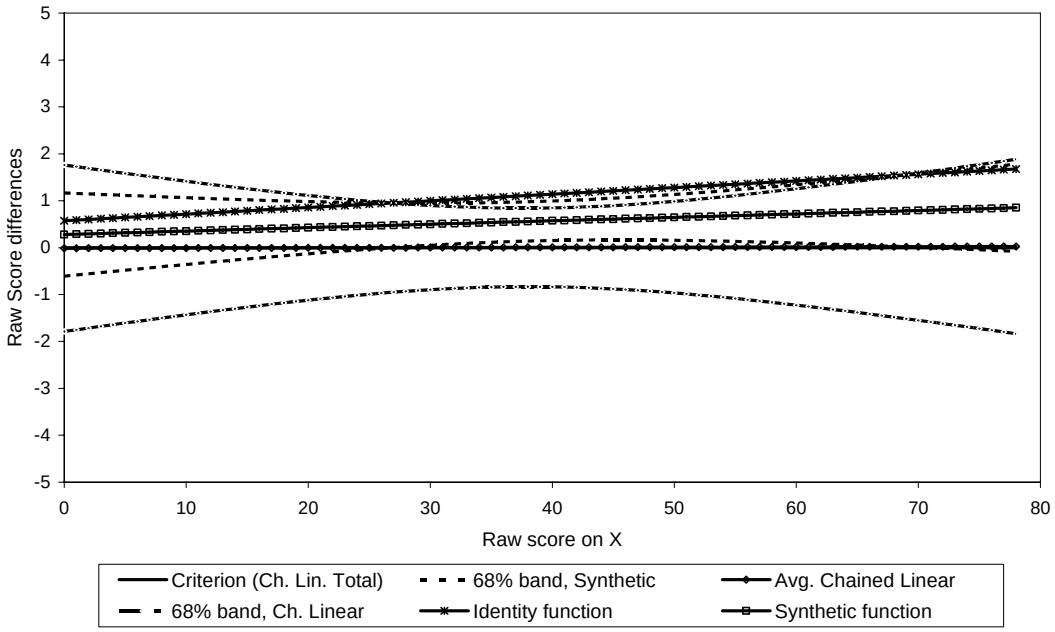

Figure 5. **Equating bias of different methods over 1,000 replications (*N* = 200) in the high-reliability external anchor design.**

Study 2

Data

Data sets from two national administrations of a licensure testing program, mostly composed of low-volume tests, were used to examine the impact of small samples on equating for combinations of sample size and equating method. The administrations took place in April 2004 with Form *X* (called *P*) and in March 2004 with Form *Y* (called *Q*). The data were collected following the NEAT design with an internal anchor as shown in Table 1.

The data sets consist of the raw sample frequencies of right scores for two nonparallel, 110-item tests[4] with 36 internal anchor items given to two samples (*P* and *Q*) from a national population of examinees (*T*). The number of examinees was 6,019 in *P* and 6,386 in *Q*. Descriptive statistics for these groups are summarized in Table 5.

As presented in Table 5, the mean of the anchor test *V* is 26.75 (±0.05) in population *P* and 26.56 (±0.05) in population *Q*, where 0.05 is the standard error of the mean. The difference between these two means (.19) indicates a negligible difference between the two populations for this type of testing program. Thus, population *Q* is as able as population *P*, as measured by *V*. In addition, the anchor test shows relatively lower reliability than the total test because in this case the anchor is a shorter version of the total test.

Table 5

Summary Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 2

	N	μ	σ	SEM	Reliability	ρ
X	6,019	78.58	10.78	4.2	.82 – .86	.88
V_P		26.75	4.08	2.3	.67	
Y	6,386	79.44	10.85	4.2	.82 – .86	.88
V_Q		26.56	4.06	2.3	.67	

The summary statistics for *X*, *Y*, and *V* in *P* and *Q* in the new samples (10, 25, 50, 100, and 200) are presented in Table 6, along with statistics for the total group. The equating procedures and statistical criterion are the same as in Study 1.

Table 6

Summary Statistics for the NEAT Design With a Low Reliability Internal Anchor for Five Different Sample Sizes

	μ_{XP}	σ_{XP}	μ_{VP}	σ_{VP}	μ_{YQ}	σ_{YQ}	μ_{VQ}	σ_{VQ}	ρ_{xv}	ρ_{vy}
Total	78.58	10.78	26.75	4.08	79.44	10.85	26.56	4.06	.88	.88
					$N = 10$					
Average	78.55	10.27	26.76	3.90	79.38	10.51	26.53	3.93	.86	.86
SD	3.47	2.75	1.33	1.06	3.41	2.62	1.26	1.01	.11	.10
Max	89.00	19.71	30.30	7.55	89.70	19.75	30.50	8.05	.99	.99
Min	66.00	4.14	22.40	1.23	59.40	4.08	20.20	1.27	.17	.15
					$N = 25$					
Average	78.58	10.59	26.77	4.02	79.41	10.77	26.55	4.02	.88	.88
SD	2.15	1.73	.81	.66	2.11	1.60	.78	.63	.05	.05
Max	85.72	16.80	29.12	6.79	85.40	15.72	28.76	6.19	.97	.98
Min	70.88	5.80	23.92	2.37	72.28	6.17	23.76	2.18	.61	.58
					$N = 50$					
Average	78.60	10.73	26.75	4.05	79.46	10.83	26.57	4.06	.88	.88
SD	1.51	1.23	.58	.47	1.57	1.21	.59	.46	.04	.04
Max	83.90	15.31	28.56	5.76	84.38	15.31	28.40	5.82	.95	.96
Min	72.44	6.85	24.36	2.68	74.00	7.56	24.48	2.80	.71	.76

(Table continues)

Table 6 (continued)

	μ_{XP}	σ_{XP}	μ_{VP}	σ_{VP}	μ_{YQ}	σ_{YQ}	μ_{VQ}	σ_{VQ}	ρ_{xv}	ρ_{vy}
					$N = 100$					
Average	78.53	10.81	26.74	4.08	79.41	10.86	26.55	4.07	.88	.88
SD	1.07	.88	.41	.34	1.06	.83	.40	.31	.02	.02
Max	82.11	14.16	27.84	5.32	82.97	13.77	27.73	5.02	.93	.95
Min	75.16	8.33	25.32	3.09	75.78	8.65	25.30	3.16	.79	.75
					$N = 200$					
Average	78.57	10.77	26.75	4.07	79.47	10.83	26.56	4.06	.88	.88
SD	.77	.61	.29	.23	.76	.56	.28	.22	.02	.02
Max	80.91	12.74	27.53	4.77	81.60	13.23	27.33	4.98	.94	.93
Min	76.14	8.95	25.85	3.38	76.53	9.10	25.52	3.40	.80	.82

Note. μ_{XP}, σ_{XP}, μ_{VP}, and σ_{VP} are the means and the standard deviations of X and V in P. μ_{YQ}, σ_{YQ}, μ_{VQ}, and σ_{VQ} are the means and the standard deviations of Y and V in Q. ρ_{xv} is the correlation between X and V in P. ρ_{vy} is the correlation between Y and V in Q. The average, SD, max, and min indicate the mean, standard deviation, maximum value, and minimum values for each statistic (e.g., μ_{XP}, σ_{XP}, μ_{VP}, σ_{VP}, and ρ_{xv}), respectively, computed over 1,000 replications.

Results

Form *X* was equated to Form *Y* using a chained linear equating method with a sample of 6,019 examinees for Form *X* and 6,386 examinees for Form *Y*. For each raw score (0 to 108) on Form *X*, the equivalent raw score on Form *Y* was determined. This raw-to-raw conversion served as the equating criterion. To examine the effectiveness of small sample equating, the same three linking methods as in Study 1 were examined. The results from each of these methods were compared to the criterion equating function. Within each sample size/equating approach, the conditional bias, SEE, RMSE, and SRMSE, which were calculated over 1,000 replications, were averaged across the entire range of scores, as in Study 1. Table 7 presents the averaged bias and equating error for combinations of sample size and equating method, along with the RMSE and SRMSE measures.

As presented in Table 7, by design, the averaged bias (1.170) and equating error (0.00) of the identity function were constant across the five sample sizes. For the chained linear method, less equating bias and error occurred as sample size increased; however, the pattern of equating bias was somewhat inconsistent. When the sample size was as small as 10, both the identity and chained linear function showed very large bias. Although there is no random error for the identity function, this function showed the greatest amount of bias (1.170) compared to the chained linear and synthetic functions when sample sizes increased. The chained linear showed the least bias (from 1.618 to .056), but it showed the greatest amount of equating error across sample sizes, 10 (11.373) to 200 (1.966). As in Study 1, the synthetic function showed the next highest level of bias (from 1.164 to .559), but it provided much smaller equating error (5.686 to .983) than did the chained linear.

As a function of equating bias and error, RMSE and SRMSE also showed a similar pattern across the five sample sizes as in Study 1. RMSE and SRMSE were smaller for both the chained linear and synthetic functions as the sample size increased. The identity showed the least RMSE for sample sizes 10 to 100, and the synthetic function showed the least RMSE when the sample size was as large as 200. Overall, the identity function seems better than any linking function in this case. The identity function showed a SRMSE of equating less than .11. The SRMSE for the synthetic function was approximately .2 only for samples as small as 50, but the chained linear equating approached this level for samples of 200 (or greater). At samples of 200, all the methods resulted in SRMSE values less than .2. The SRMSE values were about twice as large as those in Study 1, reflecting the differences in the reliabilities, correlations, and nature of the two assessments.

Table 7

Summary of Differences Among Various Equating Functions Using Small Samples in a NEAT Design With a Low Reliability Internal Anchor

Deviance measures	Chained linear–criterion	Identity–criterion	Synthetic–criterion
		$N = 10$	
Bias	1.618	1.170	1.164
SEE	11.373	.000	5.686
RMSE	11.487	1.170	5.804
SRMSE	1.059	.108	.535
		$N = 25$	
Bias	.769	1.170	.809
SEE	6.235	.000	3.118
RMSE	6.283	1.170	3.221
SRMSE	.579	.108	.297
		$N = 50$	
Bias	.190	1.170	.646
SEE	4.300	.000	2.150
RMSE	4.305	1.170	2.245
SRMSE	.397	.108	.207
		$N = 100$	
Bias	.056	1.170	.588
SEE	2.984	.000	1.492
RMSE	2.985	1.170	1.604
SRMSE	.275	.108	.148
		$N = 200$	
Bias	.096	1.170	.559
SEE	1.966	.000	.983
RMSE	1.968	1.170	1.131
SRMSE	.181	.108	.104

Note. These values for the bias, SEE, RMSE, and SRMSE were averaged across the entire range of raw scores that were calculated over 1,000 replications.

Figures 6 to 10 plot the conditional bias of equating for each method, along with an error band representing plus or minus one empirical conditional SEE. As in Study 1, this range represents the 68% confidence interval for the chained linear equating function. Even with a sample size of 200, the 68% band for chained linear equating was very wide, implying severe fluctuation across 1,000 replications. Although the 68% band was very wide for raw scores less than 70, it became narrower as sample size increased for raw scores of 70 to 100. The chained linear function seems to outperform the identity function at a certain score region when the sample sizes were as large as 100.

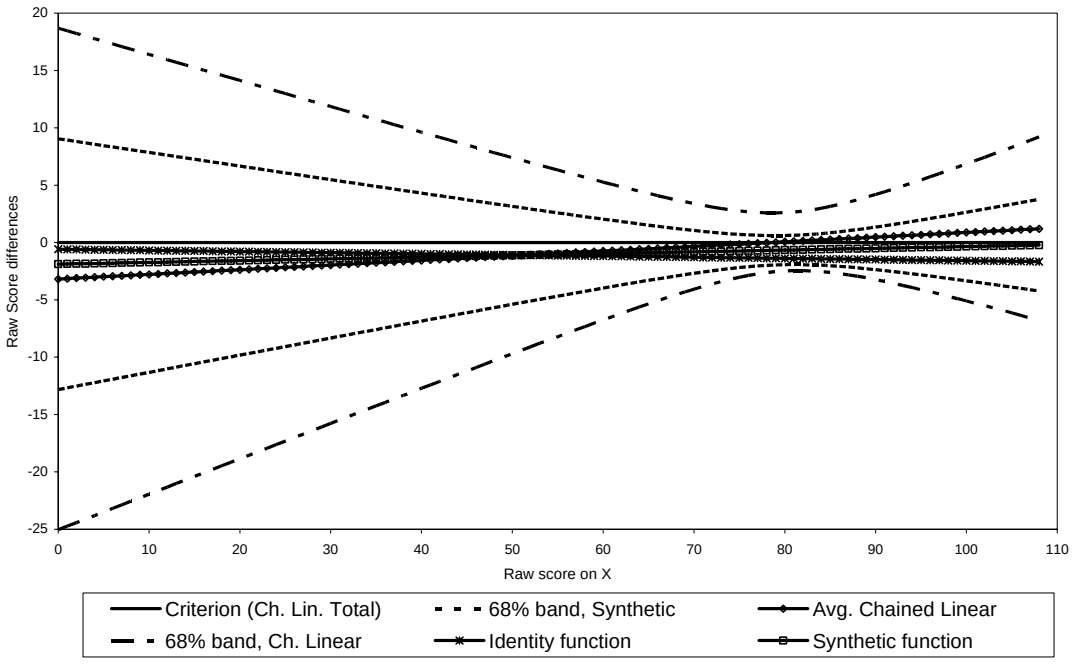

Figure 6. **Equating bias of different methods over 1,000 replications (*N* = 10) in the low-reliability internal anchor design.**

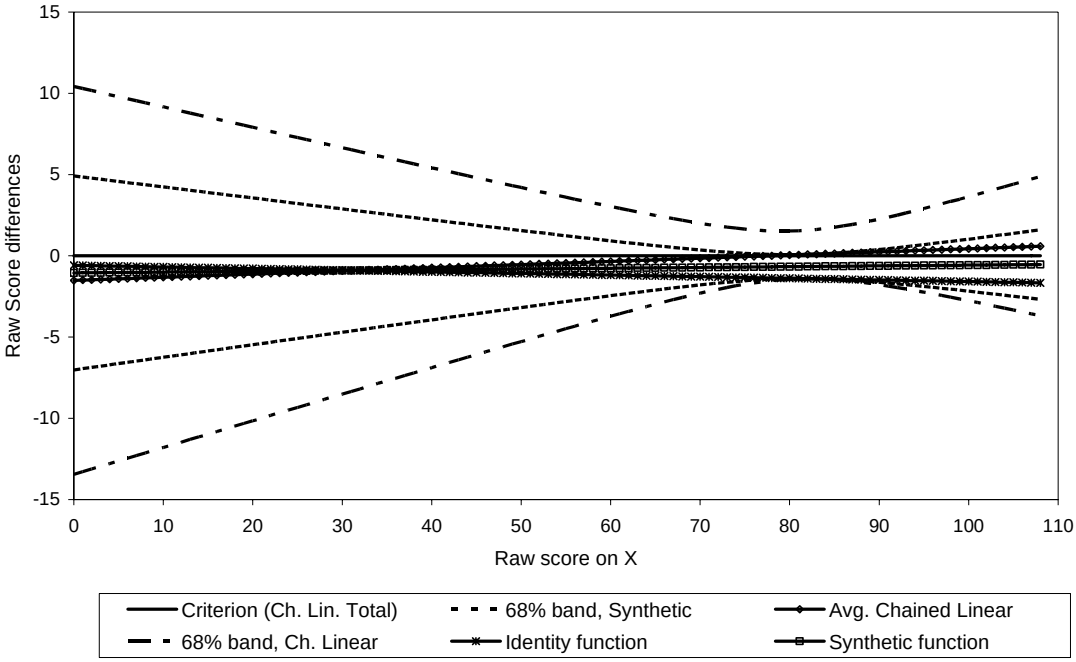

Figure 7. **Equating bias of different methods over 1,000 replications (*N* = 25) in the low-reliability internal anchor design.**

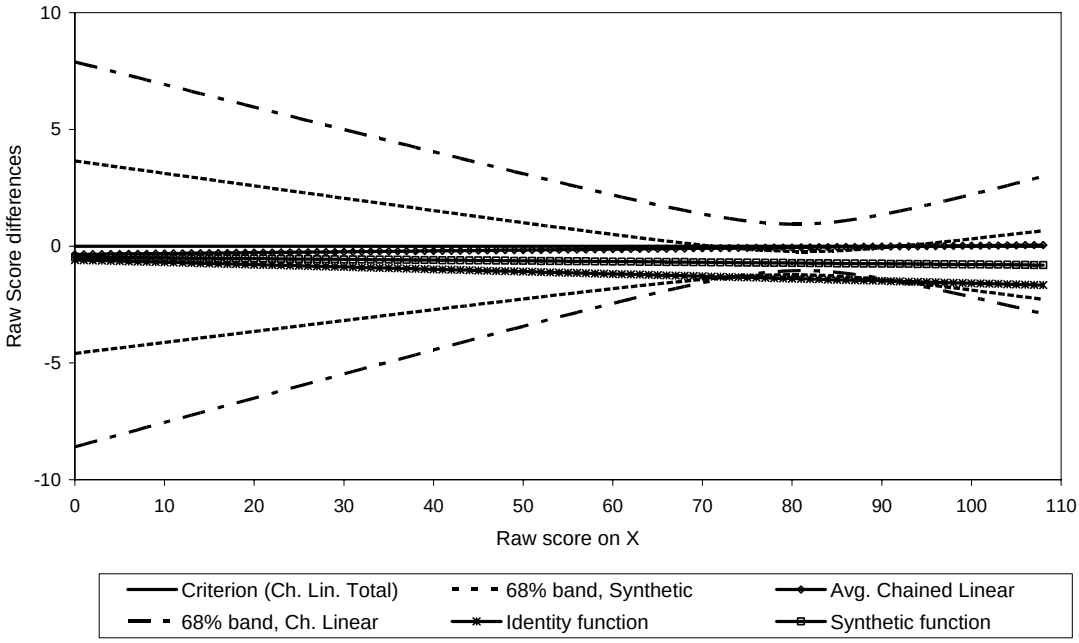

Figure 8. **Equating bias of different methods over 1,000 replications (*N* = 50) in the low-reliability internal anchor design.**

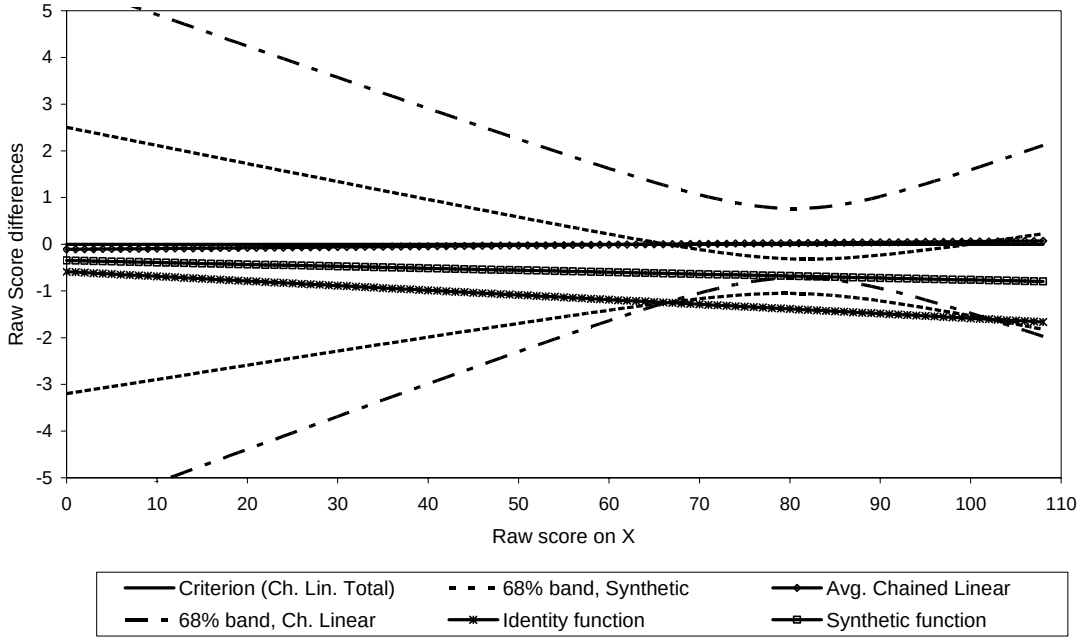

Figure 9. **Equating bias of different methods over 1,000 replications (*N* = 100) in the low-reliability internal anchor design.**

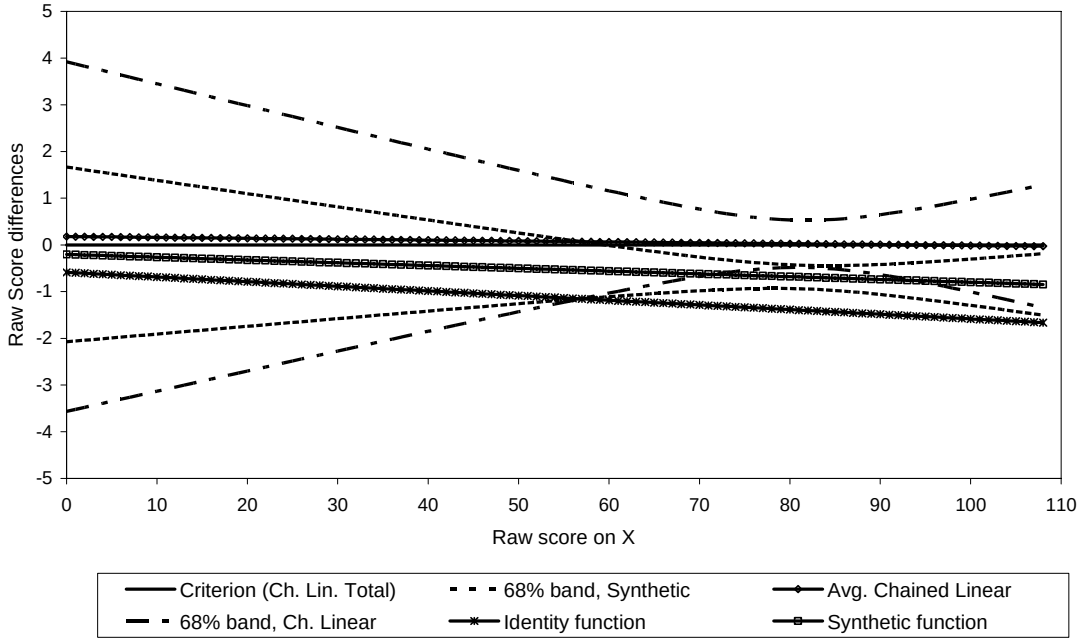

Figure 10. **Equating bias of different methods over 1,000 replications (*N* = 200) in the low reliability internal anchor design.**

As in Study 1, when sample sizes were small (less than 200), the synthetic function outperformed the chained linear function. The identity function was preferable to either alternative for samples as small as 25. Again, the synthetic function showed smaller equating error for samples as large as 200 compared to the chained linear function. The chained linear function, however, performed as well as the synthetic function when the sample size was 200. As in Study 1, overall RMSE/SRMSE for the synthetic function were smaller than those for the identity function when sample sizes were greater than 100. When only the cut-score region was considered, however, RMSE/SRMSE for the synthetic function were much smaller than those for the identity function even when the sample was as small as 50. As presented in Table 5, the means of the internal anchor tests were almost the same across samples *P* and *Q*, indicating equal abilities for the two groups. The means for the total tests were also the same, indicating equal difficulty for the two test forms. Thus, the use of the identity function might be useful in this case, along with some form of equating.

As presented in Figure 8, bias of the identity function was outside of the 68% band of the chained linear function, particularly for raw scores greater than 65. This indicates that some nontrivial linking function may be necessary for sample sizes as large as 50 to enhance the accuracy of examinees' pass-fail designations. The benefits of equating were clear when the sample size was as large as 100. As expected, the synthetic function showed a narrower error band than did the chained linear across samples sizes. The chained linear function seems to achieve a level similar to that of the synthetic function when samples sizes are close to 100. With samples of less than 100, however, the 68% band for the chained linear function was still wide for the cutoff score range (raw scores of 64 to 77). In Study 2, the tests employed had different passing scores or performance (i.e., proficiency) standards for different states. In this case, equating accuracy is much more important in the passing score range than at other score points. For assessments in which a pass/fail decision is rendered based upon the total test score, the focus should be on the standard errors of equating at and near the cutoff score. Accordingly, one way to assess small-sample equating results is to examine their impact on examinees with respect to pass/fail decisions, as Skaggs (2005) discussed. For that reason, summary measures were calculated only for the cutoff score region (raw score points from 61 to 80). These results are presented in Table 8.

Table 8

Summary of Differences Among Various Equating Functions Using Small Samples in a NEAT Design With a Low Reliability Internal Anchor: Only for the Cutoff Score Region

Deviance measures	Chained linear–criterion	Identity–criterion	Synthetic–criterion
		$N = 10$	
Bias	.399	1.293	.812
SEE	3.958	.000	1.979
RMSE	3.978	1.293	2.139
SRMSE	.367	.119	.197
		$N = 25$	
Bias	.181	1.293	.717
SEE	2.266	.000	1.133
RMSE	2.273	1.293	1.341
SRMSE	.210	.119	.124
		$N = 50$	
Bias	.091	1.293	.690
SEE	1.549	.000	.775
RMSE	1.552	1.293	1.037
SRMSE	.143	.119	.096
		$N = 100$	
Bias	.012	1.293	.642
SEE	1.106	.000	.553
RMSE	1.106	1.293	.848
SRMSE	.102	.119	.078
		$N = 200$	
Bias	.044	1.293	.625
SEE	.753	.000	.376
RMSE	.754	1.293	.730
SRMSE	.069	.119	.067

Note. The values for the bias, SEE, RMSE, and SRMSE were averaged across the range of cutoff scores that were calculated over 1,000 replications.

As presented in Table 8, again, across all sample sizes the identity function showed the greatest amount of linking bias (1.293) compared to the chained linear and synthetic functions. The chained linear showed the least bias (from .399 to .012), and the synthetic function showed the next highest level of bias (from .812 to .625). Although the chained linear still showed larger linking error across sample sizes 10 (3.958) to 200 (.753) than did the others, its linking error was much smaller when only the cutscore region was considered. Consequently, RMSE and SRMSE were also much smaller for the chained linear and synthetic functions. When equating samples are as small as 25, the identity function seems preferable relative to the other two functions. When sample sizes increase, another linking function may be necessary to enhance the accuracy of pass/fail designations. As a result, the synthetic function showed the least RMSE/SRMSE when samples are as large as 50. The benefit of the identity function (large bias but no equating error) was not clear when sample size was close to 100.

Discussion

One requirement in equating is to have a large enough sample to produce stable and accurate results. Many testing programs (e.g., certification tests), however, are low volume in nature; therefore, it is often hard to obtain as many as 50 examinees for test equating. In the present study, we introduced an approach, called the *synthetic linking function*, to conduct test linking with small samples. Essentially, the synthetic function uses a weighting system to form a compromise between the identity function and the sample equating. We compared the linking results of the identity, the synthetic, and the chained linear functions with the linking criterion with respect to both random and systematic linking errors. Specifically, we examined the linking bias and linking error for two testing programs; one program used a highly reliable external anchor (Study 1), whereas the other used an internal anchor with a lower reliability (Study 2).

We maintain that when equating samples are neither representative nor large the identity or the synthetic function may be preferable to a traditional equating function. An equating function used with very small samples may be more harmful than the identity function due to the effect of sample size on random linking error. As a result, small-sample equating may not accurately represent the true equating relationship. The present findings converge with previous recommendations (Harris, 1993; Kolen & Brennan, 2004) regarding minimum sample sizes. The chained linear function, even with the largest sample size in this study (200), produced substantial equating error, revealing inferiority to the identity function. As long as two test forms

are well-designed and almost parallel, the identity function is likely to do less harm to examinees than conventional equating. It should be noted, however, that use of the identity function might be inappropriate when the difference between the test forms and the differences in the shape of their respective score distributions are both substantial.

The results of the present study were very consistent, thereby leading to similar conclusions for small sample equating. The two studies indicated that the proposed synthetic function method might be an alternative when sample sizes are small and groups differ in ability. In both studies, the synthetic function, and even the identity, outperform the chained linear method based on very small samples. Even with samples as large as 200, the synthetic function method was preferred to the chained linear equating method regardless of anchor quality. The chained linear function showed the greatest amount of linking error, although its bias was relatively small. In addition, in both studies, RMSE/SRMSEs for the synthetic function were smaller than those for the identity function when sample sizes were greater than 100. The synthetic function exhibited lower linking error at the expense of a small amount of bias. It is worth noting that when the sample size increases, an alternative to the identity function is normally needed to reduce the total error of test linking.

The two studies presented here were conducted on two assessment programs having inherently different data characteristics. In both studies, after examining data and testing program characteristics, we applied the same weights when synthesizing the chained linear equating function with the identity function. Study 2 indicates that the identity function is not necessarily more biased for tests with relatively low reliability and low correlations with the anchor as long as the groups are similar in ability and the tests are similar in difficulty. However, for all sample sizes in Study 2, the SRMSE values are much larger, more than twice as large, as in Study 1. One reason might be the relatively smaller examinee groups available for the linking criterion in Study 2 (N_X = 6,019; N_Y = 6,386) compared to Study 1 (N_X = 10,634; N_Y = 11,321). This difference may have affected computation of the equating criterion. Another reason may be that both reliabilities and correlations between the tests and the anchor scores are much lower in Study 2 than in Study 1.

The investigation of equating accuracy may differ according to the specific properties of each testing program. In Study 2, the tests employed had different passing scores or performance (i.e., proficiency) standards for different states. In this case, linking accuracy is much more

important in the passing score range than at other score points. Based on empirical findings, Parshall et al. (1995) stated that if the cutoff score is near the mean score for the test, where standard errors are smallest, equating can probably be supported for small samples. Conversely, when cutoff scores are not near the mean, the magnitudes of the standard errors in the region of the cutoff score may make the small sample result less tolerable. This assertion was partially supported in our Study 2; when we focused solely on the cutoff score region, the identity function was preferable with very small samples (e.g., 10 or 25), but the use of the synthetic function was more promising than other functions with moderately small samples (e.g., 50 or 100). This implies that some nontrivial linking function is preferable when sample sizes increase.

Limitations and Future Research

Some cautions should be mentioned with respect to properties of the synthetic linking function. In the present study, we did not thoroughly investigate its properties with respect to particular requirements (e.g., equal construct, equal reliability, equity, population invariance, and symmetry) that are often regarded as basic to all test equating (Dorans & Holland, 2000). In general, the symmetric property is required for a relationship to be considered an equating relationship. However, the synthetic function can meet this requirement when the ratio of total score variance ($\sigma_{XP} / \sigma_{VP}$) to anchor score variance in P is exactly the same as the ratio ($\sigma_{YQ} / \sigma_{VQ}$) in Q. In this case, the chained linear equating function will have the same slope (i.e., 1) as the identity function, as Equation 3 shows. Under this condition, any weight system (e.g., .5/.5; .2/.8; .7/.3) performs well without violating the symmetric property.

The fact that the synthetic function can meet the symmetric property only under certain conditions may make the synthetic function less attractive. In practice, it may be difficult to obtain the same ratio across both groups ($[\sigma_{XP} / \sigma_{VP}]=[\sigma_{YQ} / \sigma_{VQ}]$), particularly in small samples, but the synthetic function may still be the method of choice in such cases.[5] However, this limitation can be eliminated when averaging any *linear* (not nonlinear) equating function with the identity; a certain transformation in the weight systems (i.e., w in Equation 5) can make the synthetic function symmetric when the slope of the linear equating function is not the same as the identity function (implying the nonsymmetry of the synthetic function). The symmetric property for the synthetic function is explained in detail in the appendix.

In addition, there are limitations in generalizing the findings of the current study in practice. Because the study examined (a) synthesis of only the chained linear equating function with the identity function and (b) synthesis of the two functions only when both were equally weighted, two issues arise: One issue involves determining which equating function should be combined with the identity function to create the synthetic function. Many other methods, such as the Levine, Tucker, or equipercentile methods, can be synthesized with the identity function. Although mean equating was not examined in the current study, it is definitely an alternative with very small samples. Of course, the selection of a certain equating function may depend on data quality (such as reliability or score distributions) and on specific equating situations. The second issue is related to the weight system for synthesizing two different functions. Although equal weighting was used here for illustration, various weight systems can be employed based on psychometric properties of the tests being equated, sample sizes of each group, test reliability, and test specification.

An objective tool to guide weighting is still lacking. One possibility is to collect empirical information from previous administrations or different forms of the same test; this information can be called a priori. The a priori information can then be used to weight the two functions. Future studies will be needed to establish a procedure for defining a priori information using historical test information.

In both studies, small samples were randomly drawn from large samples. There was no attempt to draw or construct unrepresentative small samples within each of the two original groups. Future research might be designed to distort the distributions of the two tests in the small samples by selecting, for example, a sample of only 10 highly able test takers. In general, more research is needed (a) to explore performance of the synthetic function using data from various testing situations across different administrations, (b) to resolve a number of issues (e.g., equating requirements, selection for the weight system), and (c) to expand the use of the synthetic function in various scenarios.

Despite the limitations of the current study, it is clear that, in some cases, the synthetic function can provide certain empirical benefits, including reduction of bias and linking error. However, as with any other approach to small-sample linking, use of the synthetic function should be approached cautiously. It is worth noting that we do not propose the synthetic function as a solution or methodological fix to a problem that is caused by poor data collection practices.

The synthetic function can be an alternative approach for the situation in which equating samples, by nature, are small. In particular, data properties should be carefully investigated before synthesizing the equating function with the identity function.

References

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. In N. J. Dorans (Ed.), Assessing the population sensitivity of equating functions [Special issue]. *Journal of Educational Measurement, 41*(1).

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the Non-equivalent group design. *Journal of Educational and Behavioral Statistics, 30*, 313–342.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4)*,* 281–306.

Hanson, B. A., Zeng, L., & Colton, D. (1991, March). *A comparison of presmoothing and postsmoothing methods in equipercentile equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Harris, D. J. (1993, April). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

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

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

Livingston, S. A. (1993). Small sample equating with long-linear smoothing*. Journal of Educational Measurement, 30*, 23-39.

Parshall, C. G., Du Bose, P., Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37–54.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309–330.

Notes

[1] In this paper, we use the standard notation of μ and σ^2 for means and variances. The subscripts usually indicate the variable in the population.

[2] There are trivial variations on the identity that would presumably be relevant for minor form differences. For instance, simple adjustment based on score range is always possible.

[3] The right minus a quarter wrong formula scores are rounded to integers, and negative scores are rounded to zero.

[4] Because two items were not scored in the X and Y forms due to content problems, the possible maximum score for the X and Y forms was 108.

[5] Because of this limitation, we have for now named this method the synthetic *linking* function (as opposed to synthetic *equating* function), as the term *linking* can be used to refer to any function used to connect the scores on one test to those on another test.

Appendix

When combining the two equating or linking functions using an ordinary weight system (called *w*), Equation A1 (Paul W. Holland, personal communication, February 15, 2006) can be used to transform the ordinary weight system into the symmetric weight system (called *W*). The symmetric weight system can symmetrize the synthetic function resulting from averaging any linear function with the identity function. For example, if $e_1(x)$ and $e_2(x)$ are two linear equating functions, such as $e_1(x) = a_1 + b_1x$ and $e_2(x) = a_2 + b_2x$, then the symmetric *w*-average of them is given by $e_w(x) = W\, e_1(x) + (1 - W)\, e_2(x)$, where the weight, *W*, is

$$W = \frac{w(1+b_2)}{1+wb_2+(1-w)b_1}. \tag{A1}$$

If the two slopes are the same ($b_1 = b_2$), then $W = w$, and the ordinary *w*-average is the symmetric *w*-average.

The synthetic function, which is a weighted average of the identity function and traditional equating function (e.g., chained linear function), can be a special case of the general averaged functions. Because the slope of the identity function is always 1 (i.e., $e_y(x) = x$), the slope of the other equating function should be 1 to maintain the symmetric property exactly. As Equation 3 indicates, the slope of the chained linear equating function will be 1 only when the following condition is met:

$$\frac{\sigma_{XP}}{\sigma_{VP}} = \frac{\sigma_{YQ}}{\sigma_{VQ}}. \tag{A2}$$

In other words, the synthetic linking function has the symmetric property when the ratio of total score variance to anchor score variance in *P* is exactly the same as the ratio in *Q*, in which case no restriction is necessary for the weight system and no transformation based on Equation A1 is needed.

In Study 1, we synthesized the chained linear equating function with the identity linking function using the same weight (i.e., .5). Therefore, one is the chained linear equating function calculated from Equation 3:

$$e_1(x) = .9852x - .557, \tag{A3}$$

and the other is the identity function:

$$e_2(x) = x. \quad \text{(A4)}$$

As can be seen, the two slopes are approximately the same; accordingly, the synthetic function used in Study 1 approximately maintains the symmetric property. When inserting the actual values from Equations A3 and A4 into Equation A1, the weight, *W*, is .5019. The weight, *W*, is close enough to all the possible values for *w*-weight, including *w* = .5 in this case. This means that under this condition any weight systems can be used to average the two functions, and the symmetric property is approximately maintained. Again, this can be confirmed simply by using the variance ratios. As presented in Table 2, the ratio of total score variance to anchor score variance in *P* (2.07 [=17.23/8.33]) is the same as the ratio in *Q* (2.04 [= 16.73/8.21]); the symmetric property is thus approximately met in the synthetic function.

The same observation was made in Study 2. The chained linear function is

$$e_1(x) = 1.01145x+.468, \quad \text{(A5)}$$

and the identity function is

$$e_2(x) = x. \quad \text{(A6)}$$

When inserting the actual slopes from Equations A5 and A6 into Equation A1, the weight, *W*, is .4986. Again, the weight, *W*, is close enough to all the possible values for *w*-weight, including *w* = .5. The ratio of total score variance to anchor score variance was 2.64 (=10.78/4.08) in *P* and 2.67 (= 10.85/4.06) in *Q*. Accordingly, the symmetric property was approximately maintained.