A Comparison of Bivariate Smoothing Methods in Common-Item Equipercentile Equating

Bradley A. Hanson
American College Testing

The effectiveness of smoothing the bivariate distributions of common and noncommon item scores in the frequency estimation method of common-item equipercentile equating was examined. The mean squared error of equating was computed for several equating methods and sample sizes, for two sets of population bivariate distributions of equating and nonequating item scores defined using data from a professional licensure exam. Eight equating methods were compared: five equipercentile methods and three linear methods. One of the equipercentile methods was unsmoothed equipercentile equating. Four methods of smoothed equipercentile (SEP) equating were considered: two based on log-linear models, one based on the four-parameter beta binomial model, and one based on the four-parameter beta compound binomial model. The three linear equating methods were the Tucker method, the Levine Equally Reliable method, and the Levine Unequally Reliable method. The results indicated that smoothed distributions produced more accurate equating functions than the unsmoothed distributions, even for the largest sample size. Tucker linear equating produced more accurate results than SEP equating when the systematic error introduced by assuming a linear equating function was small relative to the random error of the methods of SEP equating. *Index terms: common-item equating, equating, log-linear models, smoothing, strong true score models.*

The frequency estimation method of common-item equipercentile equating uses the bivariate distributions of common and noncommon item scores for the test forms to be equated (Braun & Holland, 1982). Smoothing the bivariate distributions of common and noncommon item scores before applying the frequency estimation method has been suggested as a way to improve the performance of this method (Rosenbaum & Thayer, 1987).

Livingston and Feryok (1987) investigated the effectiveness of log-linear model bivariate smoothing (Holland & Thayer, 1987, 1989; Rosenbaum & Thayer, 1987) in the frequency estimation method of common-item equipercentile equating. They concluded that smoothing could result in more accurate equating than not smoothing. A significant limitation of the Livingston and Feryok study is that the results were based on only 12 simulated samples (three for each of four sample sizes considered).

Two methods of bivariate smoothing were examined here. These methods—log-linear model bivariate smoothing and a bivariate smoothing method based on the four-parameter beta binomial model (Lord, 1965)—may be useful in frequency estimation common-item equipercentile equating. The relative performance of these two methods was investigated using two datasets. The performance of the smoothed equipercentile (SEP) methods was compared with linear methods of common-item equating (Kolen & Brennan, 1987).

Equating Methods

Equipercentile Equating

In the common-item nonequivalent groups equating design, new and old forms of a test each con-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 4, December 1991, pp. 391-408

0146-6216/91/040391-18$2.15

sist of two test parts, with one part common between the two forms. A sample of examinees from one population (Population 1) is given the new form and a sample from another population (Population 2) is given the old form (to which the new form is to be equated). Let V_p be the random variable representing the test score (which here was the number of items answered correctly) on the items common to both test forms, for a random examinee from population p. Let the random variables G_p and H_p represent the test scores on the items unique to the new and old forms of the test (the noncommon items), respectively, for a random examinee from population p. In the nonequivalent groups design, realizations of the random variables G_1, V_1, H_2, and V_2 are observed. The number of common items is denoted K_{eq}, and the number of noncommon items, assumed to be the same for both forms, is denoted K_{neq} (all the results presented are easily extended to the case in which the number of noncommon items is allowed to be different for the two forms). The total number of items for each form is denoted K_{tot} (where $K_{tot} = K_{eq} + K_{neq}$).

If the test score $X_p = G_p + V_p$ is to be equated to $Y_p = H_p + V_p$, then the part of the test common to the old and new forms is called an internal anchor. If test score G_p is to be equated to H_p, then the common test part is called an external anchor.

A mixture of Populations 1 and 2, called the synthetic population (Braun & Holland, 1982), is used to define the equipercentile equating function. It is defined by the weights w_1 and w_2 (where $w_1 + w_2 = 1$) applied to Populations 1 and 2, respectively. For example, the probability that the total score on the new form in the synthetic population is equal to i [$P(X_s = i)$] is given by $P(X_s = i) = w_1P(X_1 = i) + w_2P(X_2 = i)$.

The equipercentile equating function is determined by the cumulative distribution functions of X_s and Y_s for an internal anchor, or by the cumulative distribution functions of G_s and H_s for an external anchor. For the following description, the case of an internal anchor is considered (the case of an external anchor is given by replacing X_s with G_s and replacing Y_s with H_s). If the random variables X_s and Y_s were continuous, then the equipercentile equating function would be given by $F_{Y_s}^{-1}[F_{X_s}(x)]$ where $F_{Y_s}(y) \equiv P(Y_s < y)$ and $F_{X_s}(x) \equiv P(X_s < x)$. All percentiles of the distribution of transformed scores on the new form (transformed using the equipercentile equating function) and the distribution of scores on the old form would be equal in the synthetic population.

Because X_s and Y_s are discrete random variables, the equipercentile equating function is not defined (F_{Y_s} is a step function so that $F_{Y_s}^{-1}$ is not defined). Holland and Thayer (1989) described a method of defining an equipercentile equating function based on X_s and Y_s by using the equating function based on continuous approximations of X_s and Y_s. The most widely used continuous approximation involves spreading out the density at each discrete score point uniformly in an interval one-half point above and below the score point (Holland & Thayer, 1989; Petersen, Kolen, & Hoover, 1989). This continuous approximation can be thought of as adding the total score random variable to a continuous uniform random variable on the interval (-1/2, 1/2) that is independent of the total score random variable. This results in a continuous distribution on the interval (-1/2, K_{tot}+1/2). Based on the resulting continuous approximations of the distributions of X_s and Y_s, the equipercentile equivalent of total score i on the new form in the case of an internal anchor is given by:

$$\frac{p^*(i) - P[Y_s < u^*(i)]}{P[Y_s = u^*(i)]} + u^*(i) - .5 \quad , \tag{1}$$

where

$$p^*(i) = P(X_s < i) + .5\,P(X_s = i) \quad , \tag{2}$$

and $u^*(i)$ is the smallest integer such that $p^*(i) < P[Y_s \leq u^*(i)]$.

In the common-item nonequivalent groups design, estimates of F_{X_s} and F_{Y_s} (or F_{G_s} and F_{H_s} in the case of an external anchor) are not directly available because there are no realizations of the random variables G_2 and H_1. The assumptions used to provide estimates of F_{X_s} and F_{Y_s} in the frequency estimation method of equipercentile equating are (Braun & Holland, 1982)

$$P(G_1 = i \mid V_1 = j) = P(G_2 = i \mid V_2 = j) \quad \text{for all } i, j, \tag{3}$$

and

$$P(H_2 = i \mid V_2 = j) = P(H_1 = i \mid V_1 = j) \quad \text{for all } i, j. \tag{4}$$

With these assumptions, the distributions of the random variables needed for equipercentile equating can be written as

$$P(X_s = i) = \sum_{j=0}^{K_{eq}} P(G_1 = i - j \mid V_1 = j) P(V_s = j) \tag{5}$$

and

$$P(Y_s = i) = \sum_{j=0}^{K_{eq}} P(H_2 = i - j \mid V_2 = j) P(V_s = j) \quad , \tag{6}$$

in the case of an internal anchor, or

$$P(G_s = i) = \sum_{j=0}^{K_{eq}} P(G_1 = i \mid V_1 = j) P(V_s = j) \tag{7}$$

and

$$P(H_s = i) = \sum_{j=0}^{K_{eq}} P(H_2 = i \mid V_2 = j) P(V_s = j) \quad , \tag{8}$$

in the case of an external anchor, where

$$P(V_s = j) = w_1 P(V_1 = j) + w_2 P(V_2 = j) \quad . \tag{9}$$

The expressions given in Equations 5 through 8 indicate that calculation of the equipercentile equating function in the common-item nonequivalent groups equating design (in the case of either an internal or external anchor) requires using the bivariate distribution of equating and nonequating item scores for both groups (the bivariate distribution of G_1 and V_1 and the bivariate distribution of H_2 and V_2). Presumably, smoothing the bivariate distributions of equating and nonequating item scores has the potential to improve estimation of the distributions given in Equations 5–8 and the equipercentile equating functions based on those distributions. Two methods of smoothing the observed bivariate distributions were used here: one based on a log-linear model and the other on the four-parameter beta binomial model.

Log-Linear Model Smoothing

Rosenbaum and Thayer (1987) suggested using log-linear models to smooth the bivariate distributions needed for equipercentile equating in the common-item equating design. These log-linear models are discussed by Holland and Thayer (1987) and Haberman (1974). For the bivariate distribution of

G_1 and V_1 (or an analogous model used for H_2 and V_2), the model can be written as

$$\log[P(G_1 = i, V_1 = j)] = \beta_0 + \sum_{k=1}^{m_{eq}} \beta_k j^k + \sum_{k=m_{eq}+1}^{m_{eq}+m_{neq}} \beta_k i^{k-m_{eq}} + \sum_{k=1}^{m_{eq}} \sum_{l=1}^{m_{neq}} \theta_{kl} i^l j^k \quad , \tag{10}$$

where $m_{eq} \leq K_{eq}$, $m_{neq} \leq K_{neq}$. A specified (small) subset of the θ_{kl} are assumed to be nonzero, and the rest are assumed to be zero. Estimates of the bivariate score frequencies based on the maximum likelihood estimates of the parameters of the model given in Equation 10 have the property that the first m_{eq} moments of the marginal distribution of the equating item score and the first m_{neq} moments of the marginal distribution of the nonequating item score are identical to the moments based on the observed frequencies. For example, if $m_{eq} = 4$, then the mean, variance, skewness, and kurtosis of the fitted distribution of the equating item score will equal the values computed from the observed frequencies. In addition, all the fitted bivariate moments corresponding to nonzero values of θ_{kl} will equal the corresponding observed bivariate moments. For example, if $\theta_{11} \neq 0$ then the covariance of the equating and nonequating item scores of the fitted distribution will equal the value computed from the observed frequencies.

The model in Equation 10 is an example of a generalized linear model (GLM); therefore, procedures used to obtain maximum likelihood estimates for GLMs can be used to provide parameter estimates for the model in Equation 10 (Agresti, 1990; McCullagh & Nelder, 1989). Maximum likelihood estimation of models such as that in Equation 10 is also discussed in Holland and Thayer (1987).

Four-Parameter Beta Binomial Model Smoothing

The four-parameter beta binomial model is a strong true score model (Lord, 1965). Under the four-parameter beta binomial model, the probability that a score random variable Z (which, in the present application, may be either the equating or nonequating item score on either the new or old form) equals i ($i = 0, \ldots, K$, where K is the number of items used to calculate the score), is given by:

$$P(Z = i) = \int_{\ell}^{u} P(Z = i \mid \tau)\, g(\tau)\, d\tau \quad , \tag{11}$$

where τ is the proportion-correct true score. The true score density $g(\tau)$ is assumed to belong to the four-parameter beta class of densities [for notational simplicity, the dependence of $g(\tau)$ on a parameter vector is not denoted]. The four-parameter beta distribution is a generalization of the usual beta distribution that, in addition to the two shape parameters (α and β), has parameters for the lower (ℓ) and upper (u) limits of the distribution ($\ell \geq 0$ and $u \leq 1$). The conditional error distribution [$P(Z = i \mid \tau)$] is assumed to be either binomial (with parameters K and τ) or Lord's (1965) two-term approximation to a compound binomial distribution. When the conditional error distribution is taken to be binomial, the model is referred to as the four-parameter beta binomial (4PBB) model. When the conditional error distribution is taken to be Lord's two-term approximation to a compound binomial distribution, the model is referred to as the four-parameter beta compound binomial (4PBCB) model.

The first step in obtaining an estimate of the bivariate distribution of equating and nonequating item scores (on either the new or old form) is to obtain estimates of the parameters in the 4PBB model (or 4PBCB model) for the marginal distributions of equating and nonequating item scores. Estimates of parameters in the 4PBB and 4PBCB models were obtained from the sample data by the method of moments. The procedure used is described in detail in Hanson (1991a).

For the 4PBCB model, the reliability used to determine the two-term approximation to the com-

pound binomial distribution (i.e., the value of Lord's k; see Lord, 1965) is that estimated under a model for two test parts given by Feldt (1975). Under this model the reliability of the equating item score in Population 1 (V_1) is

$$\frac{\lambda^2\gamma}{\sigma^2_{V_1}} \quad , \tag{12}$$

and the reliability of the nonequating item score in Population 1 (G_1) is

$$\frac{(1-\lambda)^2\gamma}{\sigma^2_{G_1}} \quad , \tag{13}$$

where

$$\lambda \equiv \frac{\sigma^2_{X_1} + (\sigma^2_{V_1} - \sigma^2_{G_1})}{2\sigma^2_{X_1}} \quad , \tag{14}$$

and

$$\gamma \equiv \frac{4\sigma_{V_1G_1}}{1 - \left(\frac{\sigma^2_{V_1} - \sigma^2_{G_1}}{\sigma^2_{X_1}}\right)^2} \quad , \tag{15}$$

where $\sigma^2_{V_1}$, $\sigma^2_{G_1}$ and $\sigma^2_{X_1}$ are the variances of the random variables V_1, G_1, and X_1, respectively, and $\sigma_{V_1G_1}$ is the covariance of the random variables V_1 and G_1. Estimates of these reliabilities are obtained by substituting sample moments for population moments. The estimated reliabilities of the equating and nonequating item scores were used to estimate the value of Lord's k (Lord, 1965, pp. 265–267) that is needed for the two-term approximation to the compound binomial distribution. The estimated reliabilities used to obtain the values of Lord's k in Population 2 were obtained in a manner analogous to that described for Population 1 by substituting V_2 for V_1, H_2 for G_1, and Y_2 for X_1.

An estimate of the bivariate distribution of equating and nonequating item scores was obtained using a method presented by Lord (1965). It was assumed that the true scores of the equating and nonequating item scores were functionally related (i.e., the equating and nonequating item sets were measuring the same construct), and that conditioned on true score, the equating and nonequating item score random variables were independent. Under these assumptions, the probability that the nonequating item score random variable for Population 1 (G_1) is equal to i ($i = 0, \ldots, K_{neq}$) and the equating item score random variable for Population 1 (V_1) is equal to j ($j = 0, \ldots, K_{eq}$) can be written as

$$P(G_1 = i, V_1 = j) = \int_\ell^u P[(V_1 = j \mid \psi(\tau)]P(G_1 = i \mid \tau)\, g_1(\tau)\, d\tau \quad , \tag{16}$$

where $g_1(\tau)$ is the nonequating item true score density for Population 1 and $\psi(\tau)$ is the function that gives the equating item proportion-correct true score as a function of the proportion-correct nonequating item true score. An equation analogous to Equation 16 for the bivariate distribution of H_2 and V_2 is given by replacing G_1 with H_2, V_1 with V_2, and $g_1(\tau)$ with $g_2(\tau)$.

The integral in Equation 16 was evaluated using 64-point Gauss-Legendre quadrature (Thisted, 1988). This method of numerical integration will work well if the function to be integrated does not exhibit any severe nonpolynomial behavior in the region of integration. An indication of the adequacy of the numerical integration is how close the sum of all elements in the bivariate distribution given

by Equation 16 is to 1. Experience has indicated that, as long as the two shape parameters of the beta distribution are both greater than 1, the sum of the elements in the bivariate distribution computed using 64-point Gauss-Legendre quadrature is within 10^{-4} of 1, and usually much closer.

When one or both of the shape parameters of the beta distribution is less than 1 (so that the beta density function approaches infinity at one or both of the limits), 64-point Gauss-Legendre quadrature does not typically produce this degree of accuracy. In the cases in which only one of the shape parameters was less than 1, the interval of integration was broken into three subintervals. 64-point Gauss-Legendre quadrature was performed for each subinterval separately, and the results summed. When the beta true score density approached infinity at the lower limit, the three intervals of integration used were $[\ell, .0001(u - \ell) + \ell]$, $[.0001(u - \ell) + \ell, .01(u - \ell) + \ell]$, and $[.01(u - \ell) + \ell, u]$, where ℓ and u are the lower and upper limits of the beta density. When the beta true score density approached infinity at the upper limit, the three intervals of integration used were $[\ell, u - .01(u - \ell)]$, $[u - .01(u - \ell), u - .0001(u - \ell)]$, and $[u - .0001(u - \ell), u]$.

When both shape parameters were less than 1, the region of integration was broken into five subintervals. 64-point Gauss-Legendre quadrature was performed for each subinterval separately, and the results summed. The five intervals of integration were $[\ell, .0001(u - \ell) + \ell]$, $[.0001(u - \ell) + \ell, .01(u - \ell) + \ell]$, $[.01(u - \ell) + \ell, u - .01(u - \ell)]$, $[u - .01(u - \ell), u - .0001(u - \ell)]$, and $[u - .0001(u - \ell), u]$.

For many situations in which one or both of the shape parameters were less than 1, this procedure provides results comparable in accuracy to those obtained using a single interval when both shape parameters are greater than 1. For extreme beta distributions (e.g., when a shape parameter is near 0), this procedure can give very inaccurate results.

Method

Two pairs of population bivariate distributions of equating and nonequating item scores for new and old forms were specified. Two corresponding population equating functions were calculated, and monte carlo methods were used to estimate equating error for several equating methods and sample sizes.

Equating Methods

Eight methods of estimating the population equating functions were compared. One of the methods, the unsmoothed equipercentile (UE) equating function, uses the estimated equipercentile equating function based on the observed data. Applying the expressions in Equations 1, 2, 5, and 6 to compute the equipercentile equating function can be problematic when there are 0 frequencies in the bivariate distribution of equating and nonequating item scores. The observed bivariate distributions can be mixed with a uniform distribution to eliminate score combinations with 0 probability. If p is the probability of a particular combination of equating and nonequating item scores based on the observed data, then the modified probability (p^*) for that combination of equating and nonequating item scores is given by $p^* = .999999p + .000001t^{-1}$, where t is the number of combinations of equating and nonequating item scores. For example, if there were 15 equating items and 85 nonequating items, then $t = 86 \times 16 = 1{,}376$.

Four estimated equating functions based on smoothed bivariate distributions were considered. Two of these were based on bivariate log-linear model smoothing (using two alternate models). One of the log-linear models used was the model that produced the most accurate equating in the study by Livingston and Feryok (1987). This model—referred to here as the Log-Linear 1 model (LL1)—had $m_{eq} = 3$, $m_{neq} = 3$, and θ_{11} nonzero, using the notation in Equation 10. The other log-linear model used—referred to here as the Log-Linear 2 model (LL2)—had $m_{eq} = 4$, $m_{neq} = 4$, and θ_{11}, θ_{21}, θ_{12}

nonzero. The LL2 model was selected because it provided an adequate fit to the bivariate test score distributions used, and it fit significantly better than the LL1 model.

The two other estimated equating functions using smoothed bivariate distributions were based on the 4PBB and 4PBCB models. Due to the two-term approximation of the compound binomial distribution, negative probabilities can exist in the smoothed bivariate distribution based on the 4PBCB model. These negative probabilities are usually very small. Negative probabilities that occurred in the smoothed bivariate distributions were replaced by 0 and the probabilities were adjusted to sum to 1. The resulting bivariate distribution (containing some 0 probabilities) was mixed with a uniform distribution in the same manner as described for computing the UE equating function with mixing proportions $1 - 10^{-8}$ and 10^{-8} for the smoothed and uniform distributions, respectively. The remaining three estimated equating functions were based on linear equating methods—the Tucker (TU) method (Kolen & Brennan, 1987), the Levine Equally Reliable (LER) method (Kolen & Brennan, 1987), and the Levine Unequally Reliable (LUR) method (Angoff, 1982, section 5.4.2).

Data

Test data from an administration of a professional licensure exam were used to define population distributions of equating and nonequating item scores for the new and old test forms and to define a resulting population equating function. The data used were from three forms of a 100-item test given on three separate test dates (the score of this test is combined with a score of another separately timed test to produce the reported score). One of these forms was selected to be equated to each of the other two forms (these two forms are referred to as "old forms"). The form that was equated consisted of 70 nonequating items and two 15-item links, one to each of the old forms. The anchor items were considered to be internal. Thus, for the form equated there were 85 nonequating items and 15 equating items corresponding to each of the old forms. The two old forms each contained 85 nonequating items and a 15-item link to the form that was equated. The three forms involved were designed to be approximately parallel. In addition, the equating links were designed to be shorter versions of the full test that would be approximately parallel except for length. The weights used to define the synthetic population were $w_1 = 1$ and $w_2 = 0$.

The data used were from 38,765 examinees who took the form that was equated (referred to as the New Form), 39,150 examinees who took the first old form (Old Form 1) and 17,824 examinees who took the second old form (Old Form 2). The group that took the New Form was more similar to the group that took Old Form 1 than to the group that took Old Form 2.

Descriptive statistics for the equating item, nonequating item, and total (equating plus nonequating items) test scores for the three forms are given in Table 1. The KR-20 reliability coefficients for the total test scores were .78, .79, and .77 for the New Form, Old Form 1, and Old Form 2, respectively. The observed total test score distributions for the three forms are given in Figure 1 along with smoothed total test score distributions calculated from the smoothed bivariate distributions of equating and nonequating item scores using a model in the form of Equation 10, with $m_{eq} = 4$, $m_{neq} = 4$, and θ_{11}, θ_{21}, and θ_{12} nonzero.

Figure 2 shows the equating functions for equating the New Form to Old Form 1 as estimated using the eight equating methods, based on the examinees who took the New Form and the examinees who took Old Form 1. Very few examinees received scores in extremes of the distribution (especially the lower extreme). For instance, 99% of the examinees who took the New Form had scores between 37 and 83 (see Figure 1).

It was originally intended that the equipercentile equating functions based on the UE equating functions would be used as population equating functions. It was decided due to the rough shape

Table 1
Test Form Descriptive Statistics for All Items (K_{tot} = 100), Nonequating Items (K_{neq} = 85), and Equating Items (K_{eq} = 15) (r = Correlation of Nonequating and Equating Item Scores)

Form and Statistic	All Items	Nonequating Items	Equating Items
New Form (Link to Old Form 1, N = 38,765; r = .53)			
Mean	62.51	52.58	9.94
SD	9.45	8.07	2.22
Skewness	-.28	-.22	-.33
Kurtosis	2.97	2.94	2.91
New Form (Link to Old Form 2, N = 38,765; r = .56)			
Mean	62.51	53.71	8.80
SD	9.45	7.96	2.32
Skewness	-.28	-.32	-.10
Kurtosis	2.97	3.06	2.69
Old Form 1 (N = 39,150; r = .54)			
Mean	65.03	55.37	9.65
SD	9.38	8.04	2.17
Skewness	-.33	-.30	-.28
Kurtosis	3.04	3.01	2.92
Old Form 2 (N = 17,824; r = .53)			
Mean	60.55	52.39	8.17
SD	9.39	7.99	2.27
Skewness	-.06	-.12	.02
Kurtosis	2.96	3.02	2.75

of the UE equating functions that the fitted bivariate distributions based on the LL2 model would be used as the population distributions. The LL2 model fit the bivariate distributions for both forms very well. The Pearson goodness-of-fit χ^2 statistics were 1,186.6 and 1,126.3 for the new and old forms, respectively [with 1,364 degrees of freedom (df)—1,376 cells in the bivariate table minus 12 model parameters]. The likelihood-ratio χ^2 statistics were 757.5 and 734.5 for the old and new forms, respectively. The fit of the marginal and conditional distributions and conditional moments up to degree 4 of the nonequating item scores given equating item scores were graphically examined; these results also indicated that for both bivariate distributions the model fit the data well. The LL1 model did not provide an adequate fit to these bivariate distributions. The differences in the likelihood-ratio χ^2 statistics between the LL2 and LL1 models were 236.1 and 200.7 for the new and old forms, respectively, based on 4 df (the difference in the number of parameters between the two models).

Figure 3 shows the equating functions for equating the New Form to Old Form 2, as estimated using the eight equating methods, based on the examinees who took the New Form and those who took Old Form 1. Again, the fitted bivariate distributions based on the LL2 model were used as the population distributions. The Pearson χ^2 statistics were 852.1 and 818.1 for the new and old forms, respectively, based on 1,364 df. The likelihood-ratio χ^2 statistics were 705.5 and 670.6 for the old and new forms, respectively. The differences in the likelihood-ratio χ^2 statistics between the LL2 and LL1 models were 302.7 and 121.6 for the new and old forms, respectively, based on 4 df.

For each of the two pairs of population bivariate distributions (equating versus nonequating item scores) used for equipercentile equating, 300 samples for each of five sample sizes (100, 250, 500, 1,000, and 3,000) were drawn. For each of the 3,000 pairs of sample bivariate distributions (five

Figure 1
Total Score Distributions

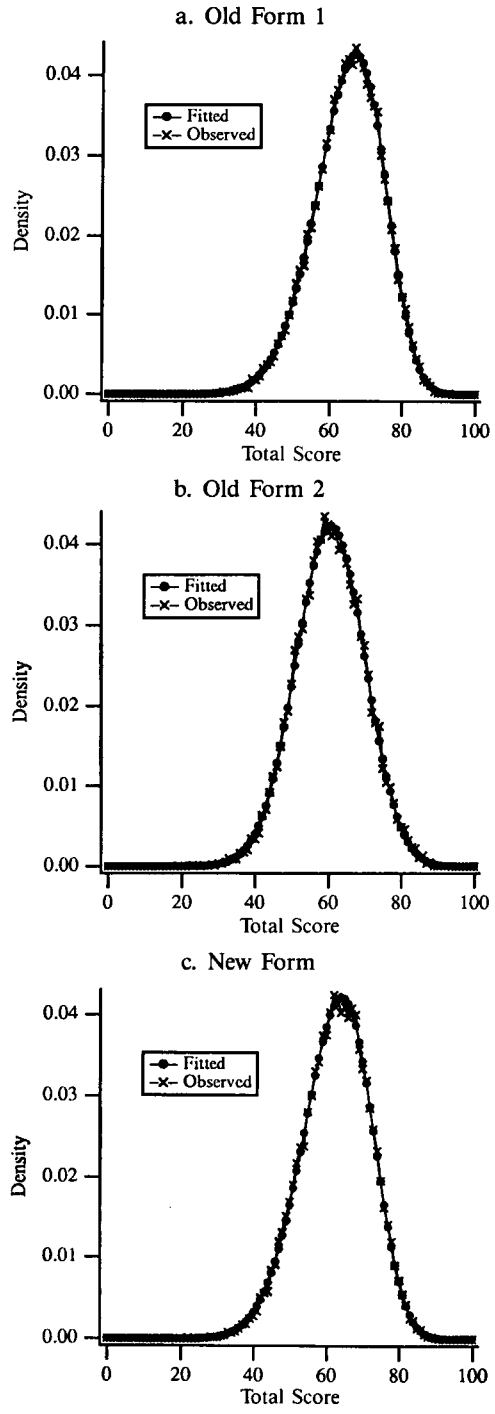
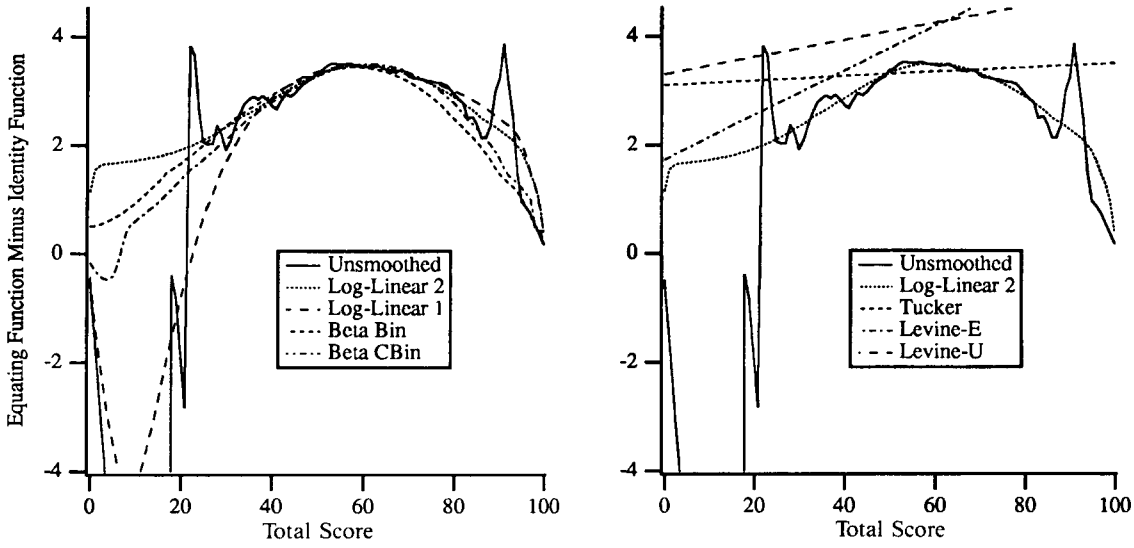


Figure 2
Equating Functions for New Form to Old Form 1 Equating Using All Data



sample sizes × two pairs of population bivariate distributions × 300 samples) the eight estimated equating functions were computed.

Criteria

If $\tilde{e}(i)$ is the estimated old form score equivalent to new form score i given by a particular equating method and $e(i)$ is the score equivalent given by the population equating function (defined using the fitted distributions based on the LL2 model), then the mean squared error (MSE) for the equating method at new form score i is given by

$$\text{MSE} = \text{E}[\tilde{e}(i) - e(i)]^2 \quad , \tag{17}$$

where E indicates the expected value, which is taken over the two pairs of random variables used to determine $\tilde{e}(i)$. The MSE can be written as

$$\text{MSE} = \text{E}[\tilde{e}(i) - \mu_{\tilde{e}(i)}]^2 + [e(i) - \mu_{\tilde{e}(i)}]^2 \quad , \tag{18}$$

where $\mu_{\tilde{e}(i)} \equiv E[\tilde{e}(i)]$. The first term in Equation 18 is the variance of $\tilde{e}(i)$, and the second term is the squared bias of $\tilde{e}(i)$.

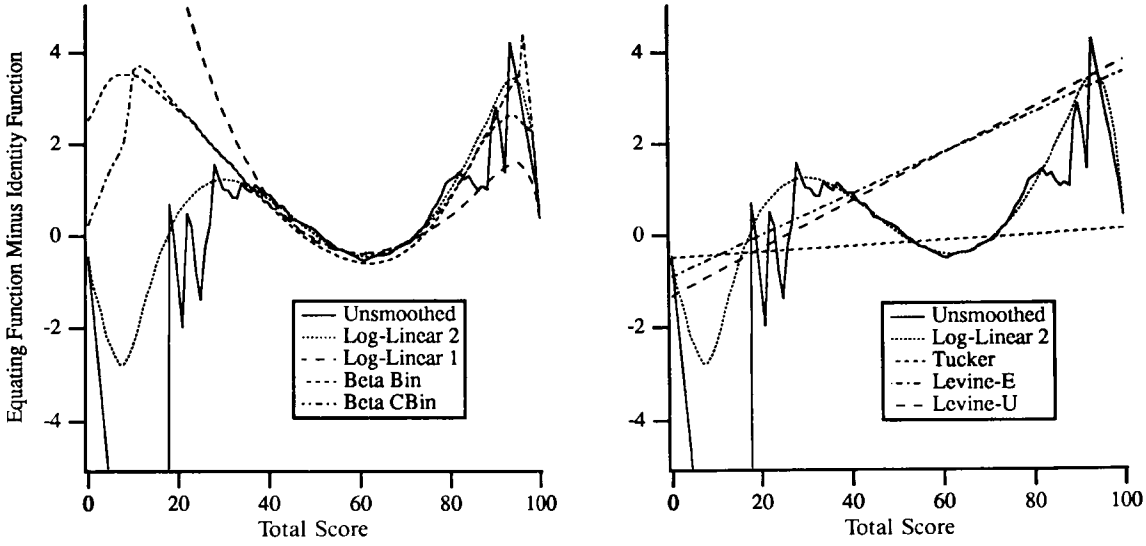
The average MSE for an equating method is given by

$$\sum_{i=0}^{K_{tot}} \text{E}[\tilde{e}(i) - e(i)]^2 P(X_s = i) \quad , \tag{19}$$

which can be written as the sum of the average variance and average squared bias:

$$\sum_{i=0}^{K_{tot}} \text{E}[\tilde{e}(i) - \mu_{\tilde{e}(i)}]^2 P(X_s = i) + \sum_{i=0}^{K_{tot}} [e(i) - \mu_{\tilde{e}(i)}]^2 P(X_s = i) \quad . \tag{20}$$

Figure 3
Equating Functions for New Form to Old Form 2 Equating Using All Data



For each of the eight equating methods for a particular sample size and pair of bivariate population distributions, the MSE at score i was estimated using the 300 pairs of simulated bivariate distributions as

$$\frac{1}{300}\sum_{s=1}^{300}[\tilde{e}_s(i) - e(i)]^2 \quad , \tag{21}$$

where $\tilde{e}_s(i)$ is the old form score equivalent of new form score i for sample s. The variance and squared bias of $\tilde{e}(i)$ were estimated in a similar manner. Estimates of the average MSE were obtained by substituting the estimates of the MSE for each score into Equation 19. An estimate of the standard error of these estimates of the average MSE [SE(MSE)] was obtained from the usual estimate of the standard error of a mean (the standard deviation divided by the square root of 300). Estimates of the average variance and average squared bias were obtained analogously using Equation 20, where $\mu\tilde{e}(i)$ was estimated by

$$\frac{1}{300}\sum_{s=1}^{300}\tilde{e}_s(i) \quad . \tag{22}$$

Results

Estimates of average squared bias, average variance, average MSE, and the standard error of the average MSE for each of the eight equating methods for the New Form to Old Form 1 equating are given in Table 2. For $N = 100$, the reported average values were based on 297, 287, and 298 simulated samples, rather than 300, for the LL2, 4PBB, and 4PBCB methods, respectively. For the LL2 method, the missing samples represent cases in which the maximum likelihood estimation did not converge. For the four-parameter beta methods, the missing samples represent cases that were dropped due to one or both of the numerical integrations (see Equation 16) not producing bivariate probability

Table 2
Average Squared Bias (SB), Variance (Var), Mean Squared Error (MSE), and SE(MSE) for Equating New Forms to Old Forms by UE, TU, LER, LUR, LL1, LL2, 4PBB, and 4PBCB Methods

Sample Size and Statistic	Equating Method							
	UE	TU	LER	LUR	LL1	LL2	4PBB	4PBCB
Equating New Form to Old Form 1								
N = 100								
SB	2.563	.057	.880	.763	.047	.041	.077	.082
Var	8.505	1.717	5.249	4.128	2.347	2.820	2.651	2.795
MSE	11.067	1.774	6.130	4.892	2.394	2.861	2.727	2.877
SE(MSE)	.553	.121	.412	.313	.229	.167	.133	.140
N = 250								
SB	.558	.056	1.147	.956	.022	.015	.035	.019
Var	3.134	.686	2.054	1.833	.839	1.022	1.133	1.067
MSE	3.693	.741	3.201	2.789	.861	1.037	1.168	1.086
SE(MSE)	.210	.041	.185	.164	.042	.046	.055	.051
N = 500								
SB	.196	.042	1.201	1.052	.006	.002	.028	.002
Var	1.352	.306	.942	.797	.375	.472	.527	.489
MSE	1.547	.348	2.143	1.849	.381	.474	.555	.491
SE(MSE)	.086	.016	.105	.090	.017	.019	.024	.020
N = 1,000								
SB	.049	.040	1.003	.900	.005	.003	.039	.007
Var	.625	.169	.492	.390	.209	.264	.288	.269
MSE	.674	.209	1.495	1.290	.214	.267	.327	.275
SE(MSE)	.040	.010	.065	.057	.011	.012	.015	.013
N = 3,000								
SB	.008	.039	1.067	.932	.003	0.000	.023	002
Var	.184	.055	.149	.123	.068	.084	.094	.080
MSE	.191	.094	1.216	1.054	.071	.084	.117	.082
SE(MSE)	.010	.003	.032	.029	.003	.004	.005	.004
Equating New Form to Old Form 2								
N = 100								
SB	2.257	.176	4.171	4.196	.039	.060	.207	.091
Var	8.487	1.609	5.570	4.473	1.955	2.692	2.832	2.932
MSE	10.744	1.784	9.741	8.669	1.994	2.752	3.039	3.023
SE(MSE)	.595	.097	.562	.515	.107	.150	.142	.153
N = 250								
SB	.469	.175	4.422	4.494	.037	.012	.122	.034
Var	2.971	.706	2.068	1.715	.852	1.161	1.270	1.216
MSE	3.440	.881	6.490	6.209	.889	1.173	1.391	1.250
SE(MSE)	.168	.042	.285	.273	.046	.057	.062	.057
N = 500								
SB	.160	.175	4.414	4.488	.038	.003	.072	.013
Var	1.332	.323	.930	.811	.387	.540	.614	.565
MSE	1.492	.498	5.344	5.299	.425	.543	.686	.578
SE(MSE)	.070	.017	.170	.171	.018	.024	.030	.025
N = 1,000								
SB	.036	.175	4.249	4.278	.033	.003	.051	.014
Var	.579	.167	.454	.372	.193	.254	.322	.280
MSE	.615	.341	4.703	4.651	.225	.257	.373	.293
SE(MSE)	.030	.009	.112	.110	.010	.012	.017	.014
N = 3,000								
SB	.005	.174	4.230	4.285	.037	0.000	.045	.008
Var	.186	.058	.147	.126	.068	.089	.117	.091
MSE	.191	.233	4.377	4.411	.104	.089	.162	.099
SE(MSE)	.008	.004	.060	.060	.004	.004	.007	.005

densities that summed to within .05 of 1. This problem is an indication that the numerical integration was not producing adequate results for these cases. Neither of these problems occurred for any of the simulated samples at any of the other sample sizes.

For the population equating function used in producing the upper portion of Table 2 in which the New Form was equated to Old Form 1, the average MSE if no equating were performed (i.e., using an identity function as the equating function) was 11.139. None of the values of average MSE reported is above this value. Thus, for this case, equating was always preferable to not equating, even for the smallest sample size.

For these data, the TU and SEP methods have much lower average MSE than the UE method for all sample sizes. The TU method has the lowest value of average MSE for all sample sizes less than 3,000. The LL1 method has lower average MSE than the other SEP methods, and has the lowest average MSE of all methods for $N = 3{,}000$. The LL2 method has lower average squared bias than the LL1 method but has higher average variance. The 4PBB method has higher average MSE than the other SEP methods for all sample sizes except 100. The average MSE for the 4PBCB method is near (in relation to the standard error) the average MSE for the LL2 method. Both Levine methods have much higher average MSE than the other methods at all sample sizes (except the UE method for N less than 500).

Estimates of average squared bias, average variance, average MSE, and an estimate of the standard error of the average MSE for each of the eight equating methods for the New Form to Old Form 2 equating are given in the lower portion of Table 2. For $N = 100$, the reported average values are based on 287 and 299 simulated samples, rather than 300, for the 4PBB and 4PBCB methods, respectively (some samples were dropped due to inaccuracies in the numerical integration).

For the population equating function used in producing these data, the average MSE if no equating were performed was .1904. Only for $N = 3{,}000$ are the values of average MSE for the equating methods less than this value. Consequently, in this case equating was only useful for very large sample sizes.

The TU and SEP methods again have lower average MSE than the UE method, with one exception. This exception is for $N = 3{,}000$ in which the UE method has lower average MSE than the TU method. The TU and LL1 methods have the lowest values of average MSE for $N = 100$ and 250. For $N = 500$ and 1,000, the LL1 method has the lowest values of average MSE. For $N = 3{,}000$, the LL2 method has the lowest value of average MSE. The 4PBB method has higher average MSE than the other SEP methods for all sample sizes. Again, both Levine methods have much higher average MSE than other methods at all sample sizes (except the UE method at $N = 100$).

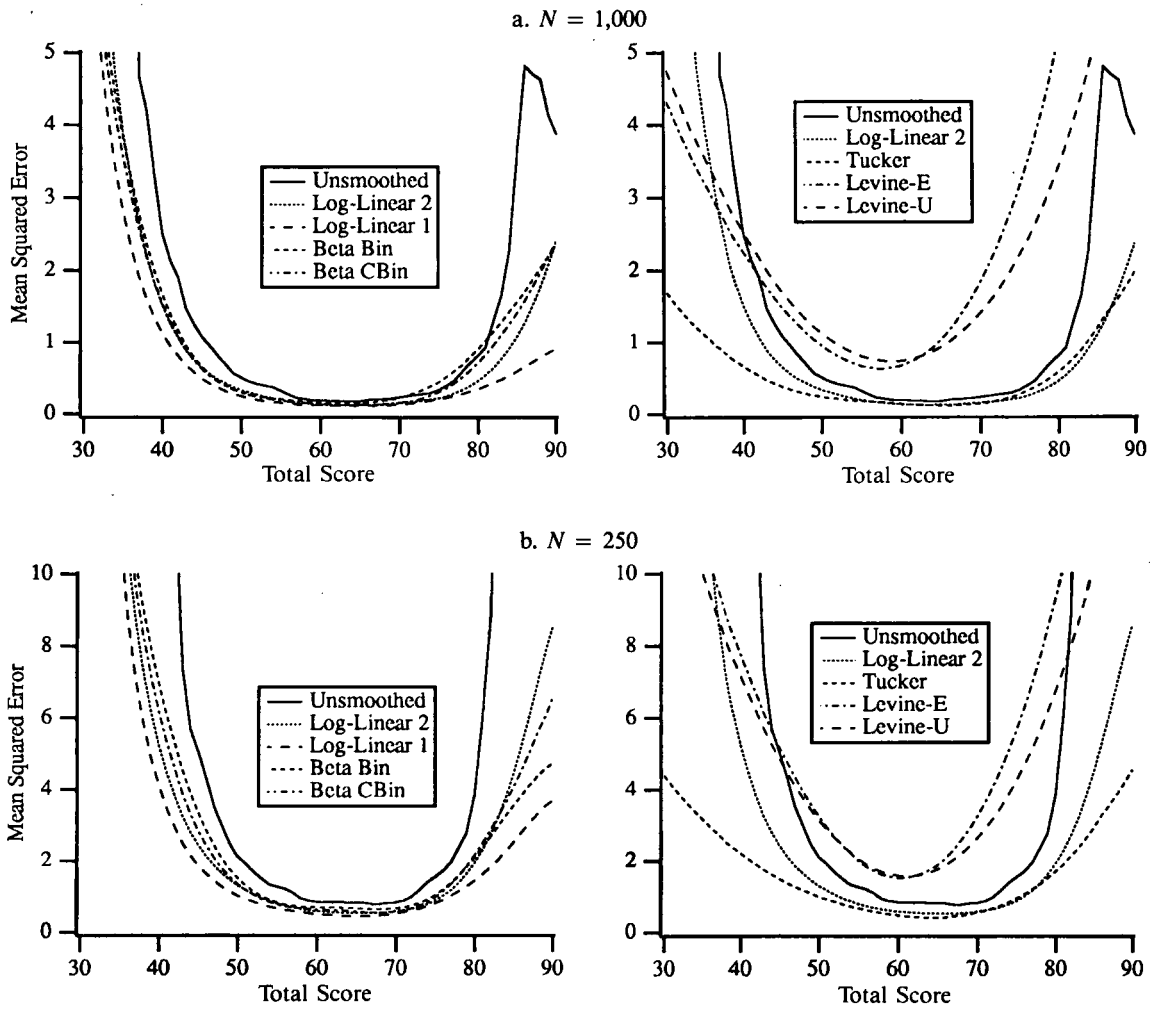
Figure 4 presents estimates of the MSE by score point (Equation 17) for the eight equating methods for the New Form to Old Form 1 equating ($N = 1{,}000$ and 250). The range of scores given in Figure 4 is 30 to 90 because there was very little probability of scores outside that range for the synthetic population (the population from which the sample that took the New Form was drawn, because $w_1 = 1$ and $w_2 = 0$). The general level of the MSE curves reflects the average MSEs given in the top portion of Table 2. Figure 4 shows that the poor performance of UE equating was largely due to the very large values of MSE for scores that were a moderate distance from the mean score.

Figure 5 shows similar results for the New Form to Old Form 2 equating. In Figure 5, the shapes of the MSE functions are different for the Levine methods than for the other methods. For $N = 1{,}000$, the highest values of MSE for the Levine methods were for scores near the mean score (which represents the bias in the Levine methods that can be seen in Figure 3).

Discussion

The results provide evidence that smoothing the bivariate distributions of equating and nonequating

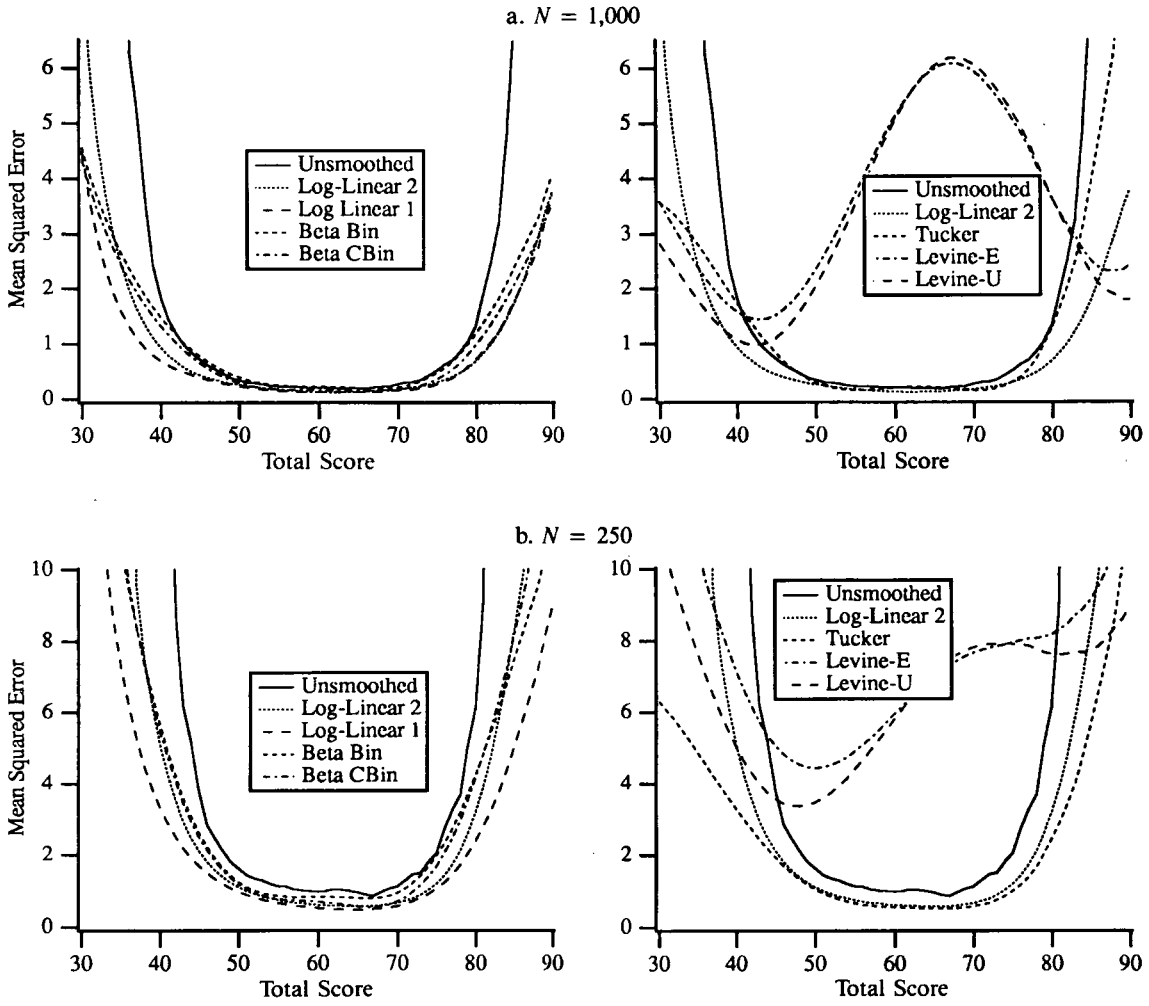
Figure 4
Mean Squared Error for Equating New Form to Old Form 1



item scores can improve estimation of the equipercentile equating function. An improvement in the estimation of the equipercentile equating function resulted from smoothing for all sample sizes considered.

The TU method worked well compared to the SEP methods when the bias introduced by the TU method was not large compared to the variance of the SEP methods. The lower average MSE of the TU method in these situations was because the variance of the TU method was less than the variance of the SEP methods. When the New Form was equated to Old Form 1, the bias of the TU method was not large and, consequently, the TU method had lower average MSE than the SEP methods for all samples sizes less than 3,000. When the New Form was equated to Old Form 2, the bias of the TU method was greater than for the Old Form 1 equating, and only for $N = 100$ did the TU method have an average MSE that was meaningfully lower (relative to the standard error) than the average

Figure 5
Mean Squared Error for Equating New Form to Old Form 2



MSE for all the SEP methods.

For bivariate smoothing based on the log-linear model, adding parameters to the model generally results in lower bias, but greater variance, of the estimated equating function. Lower average MSE may result with a model that has fewer parameters than with a model that has more parameters—even when the more complex model is the model from which the data were generated—if the bias introduced by the simpler model is small compared to the variance of the more complex model. This is especially likely to occur for small sample sizes, for which the average variance that results from a model with a large number of parameters is likely to be high. For the test forms used here, the LL1 method (with eight parameters) had lower average MSE than the LL2 method (with 12 parameters) except at the largest sample size for the New Form to Old Form 2 equating, even though the population distributions were defined using the LL2 model.

The equating method based on the 4PBCB model had lower average mean square error than the method based on the 4PBB model (with one exception at $N = 100$). This result is consistent with the results of Lord (1965) who found that bivariate test score distributions were better fit using the 4PBCB model than the 4PBB model.

The methods based on the log-linear models generally had lower values of average MSE than the method based on the 4PBCB model. In most cases, the difference in the average MSE between the LL2 and 4PBCB models was not large relative to the standard errors of the average MSE. The 4PBCB model may have been at a disadvantage relative to the log-linear models because the criterion equating function was based on the fitted distributions from the LL2 model. It is possible that the reported bias and MSE of the method based on the LL2 model were, to some extent, artificially low. Because the LL2 model fit the data used in this study very well, this effect is probably not large.

The Levine methods performed very poorly. Both Levine methods rely on assumptions concerning true score moments, and the classical congeneric weak true score model is assumed to hold for the equating and nonequating item scores (Brennan, 1990). The poor performance of the Levine methods may have been partly due to these assumptions not being well met for the data used. In addition, the LUR method is a true-score equating method that attempts to meet a different equating criterion than the observed-score equating criterion used here (Hanson, 1991b).

A major problem in attempting to provide recommendations based on these results is that values of MSE were computed only for two population equating functions based on three forms of a single test. The statistical and psychometric properties of the test forms used to define population quantities are typical of those for certification/licensure tests, with the exception that the length and reliability of the test forms were less than usual for a test of this type (recall that the test used in this study was one of two tests used in computing the score reported to an examinee). Livingston and Feryok (1987) reported that smoothing of bivariate distributions resulted in more accurate equipercentile equating in the common-item equating design using a college placement test. The results in the present study and results reported by Livingston and Feryok (1987) indicate that bivariate smoothing can be effective in common-item equating for two different types of tests. These results suggest some promise that bivariate smoothing may be effective in practice for other tests.

It is likely that an important factor in the performance of the smoothed equipercentile methods in practical situations is the appropriateness of the models used for smoothing. In using any of the smoothed equipercentile methods, the fit of the model used for smoothing to the bivariate distributions of equating and nonequating item scores should be evaluated. An important condition for using a smoothed equipercentile method in practice would be that the model used for smoothing fit the data fairly well. Assessment of model fit may involve formal tests of model fit (χ^2 goodness-of-fit statistics), and informal analyses of model fit such as residual analyses and various graphical displays (e.g., fitted versus observed frequencies for marginal distributions of common and noncommon item scores; fitted versus observed frequencies for conditional distributions of noncommon item score conditioned on common item score; or fitted versus observed conditional mean, standard deviation, skewness, and kurtosis of noncommon item scores as a function of common item score).

The results of this study and practical experience with data from several testing programs indicate that the log-linear model and the four-parameter beta compound binomial model often provide an adequate fit to observed bivariate distributions of test scores. The log-linear model has an advantage over the four-parameter beta compound binomial model in that it can potentially fit a wider class of bivariate distributions. A cost involved in this greater flexibility is that a model selection process must be used to select a particular log-linear model to use. In the present research, two fixed log-linear models were used. In applied settings, the user would likely evaluate several log-linear models

and select the simplest model that fits the data adequately. Haberman (1974) discussed model selection for models such as those given in Equation 10. Agresti (1990, Chapter 7) discussed some general methods for selecting log-linear models.

The process of selecting a log-linear model could introduce errors that were not included in the present simulations. For example, Hanson (1990) compared smoothing of univariate test score distributions based on the four-parameter beta binomial model and a log-linear model analogous to the model of Equation 10 for univariate distributions. Hanson used a model selection process for each simulated sample to select a log-linear model to use, and found that the four-parameter beta binomial model provided more accurate results than the log-linear model for all sample sizes less than 5,000. The log-linear model would likely have performed better in Hanson (1990) if a fixed model had been used for all samples, as in the present research. Conversely, the log-linear model in the present study might have been less accurate if some model selection procedure had been used for each sample to select a model.

The results suggest that another important factor in the performance of the methods in practical situations is the sample sizes available for equating. For very large sample sizes, smoothing the bivariate distribution of equating versus nonequating items may result in more equating error than not smoothing (due to the bias introduced by smoothing). For small sample sizes, smoothing may be effective, but Tucker linear equating, with lower variance than smoothed equipercentile equating, may result in less equating error. For very small sample sizes, it is likely that in many cases not equating would result in lower average MSE than equating using any method of equating (i.e., the bias that results by not equating would be less than the random error that would be introduced by equating).

Kolen and Jarjoura (1987) present a postsmoothing method of equipercentile equating using cubic splines for the common-item nonequivalent groups design. Future research should compare the performance of the postsmoothing method using cubic splines to the presmoothing methods studied here.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–69). New York: Academic Press.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Brennan, R. L. (1990). *Congeneric models and Levine's linear equating procedures* (Research Rep. No. 90-12). Iowa City IA: American College Testing.

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika, 40,* 557–561.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30,* 589–600.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions* (Research Rep. No. 90-4). Iowa City IA: American College Testing.

Hanson, B. A. (1991a). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Research Rep. No. 91-5). Iowa City IA: American College Testing.

Hanson B. A. (1991b). A note on Levine's formula for equating unequally reliable tests using data from the common-item nonequivalent groups design. *Journal of Educational Statistics, 16,* 93–100.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Rep. No. 87-31). Princeton NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Research Rep. No. 89-7). Princeton NJ: Educational Testing Service.

Kolen, M. J., & Brennan R. L. (1987). Linear equating models for the common-item nonequivalent population design. *Applied Psychological Measurement, 11,* 263–277.

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design.

Psychometrika, 52, 43–59.

Livingston, S. A., & Feryok, N. J. (1987). *Univariate versus bivariate smoothing in frequency estimation equating* (Research Rep. No. 87-36). Princeton NJ: Educational Testing Service.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30,* 239–270.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall.

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

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40,* 43–49.

Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation.* New York: Chapman and Hall.

Author's Address

Send requests for reprints or further information to Brad Hanson, American College Testing, P. O. Box 168, Iowa City IA 52243, U.S.A. The source code for C functions that compute the equating function estimates discussed in this paper is available from the author.