Standard Errors of Tucker Equating

Michael J. Kolen
The American College Testing Program

Large sample standard errors are derived for the Tucker linear test score equating method under the common item nonequivalent populations design. Standard errors are derived without the normality assumption that is commonly made in the derivation of standard errors of linear equating. The behavior of the standard errors is studied using a computer simulation and a real data example. In the simulation, the derived standard errors were reasonably accurate. In the real data example, the derived standard errors agreed closely with standard errors estimated using Efron's (1982) bootstrap.

Test form equating of observed scores adjusts for small differences in difficulty among multiple forms of a test for a specified population of examinees. Such equating requires a design for collecting data and a method for equating forms. The *common item nonequivalent populations design* is a design in which two groups of examinees from different populations are each administered different test forms that have a subset of items in common (Angoff, 1971, pp. 579–583). For this design, only one form of a test needs to be administered on a given test date, and this is one reason for its popularity. The *Tucker linear method,* which is used extensively for equating under this design, is examined in the present paper. Angoff (p. 579) recommended the Tucker method when the examinee groups used to conduct the equating are not widely different in ability.

Standard errors of equating are a means for expressing the amount of error in test form equating that is due to examinee sampling. For a given score on one form of a test, the error in estimating its equated score on another form is often indexed by a standard error. These standard errors generally differ by score level. Standard errors of equating are used (1) as a means for expressing equating error when scores are reported, (2) in the estimation of sample size required to achieve a given level of equating precision, and (3) as a basis for comparing equating methods and designs.

The primary purpose of this paper is to derive large sample standard errors for the Tucker method of linear equating and to evaluate their accuracy. Standard errors were derived under the commonly made normality assumption as well as under less restrictive assumptions. The derived expressions were evaluated using a computer simulation and a real data example based on test forms from a professional certification

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 2, June 1985, pp. 209–223

0146-6216/85/020209-15$2.00

testing program. In the real data example, standard error estimates based on the derived expressions were compared to standard errors of Tucker equating estimated using the bootstrap method described by Efron (1982).

When the Tucker method is applied, the examinee groups taking each of the test forms to be equated are weighted to form a combined group. To accommodate any of the weighting schemes that have been considered in the literature, the Tucker method is presented in a general form that allows the investigator to choose the weights. Presentation of the Tucker method and its standard errors in such a general form necessitates a detailed development of the method in the present paper.

Tucker Common Item Equating With Nonequivalent Populations

Multiple forms of a test to be equated are designed to be similar in content and statistical characteristics. For the common item nonequivalent populations design, a new form is equated to an old (previously equated) form using a set of items that are common to the two forms. The set of common items is constructed to be similar to each of the full length forms in content balance and in the statistical characteristics of its items. Scores on the common items may contribute to the total score on each form (an internal set of common items) or they may not contribute (an external set of common items). The new and old forms are administered to examinees from different populations. To accomplish observed score equating, a decision must be made on how to combine these two populations. The combined population, which has been referred to as the *synthetic population* by Braun and Holland (1982), is a weighted combination of the distribution functions of the two populations. In this way, the synthetic population can be thought of as a mixture of two populations.

Refer to the new test form as X, the old form as Y, and the set of common items as V. Examinees from Population 1 are administered X and V. Examinees from Population 2 are administered Y and V. Consider that these two populations are weighted using proportional weights w_1 and w_2 (where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$) to form the synthetic population. The linear equation for equating scores on X to the scale of Y is

$$\ell(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y) \quad . \tag{1}$$

In this equation $\mu_s(X)$, $\mu_s(Y)$, $\sigma_s(X)$, and $\sigma_s(Y)$ are the means and standard deviations of scores on X and Y for the synthetic population, and $\ell(x)$ is the value of the linear equating function at x.

The parameters in Equation 1 depend on the parameters in Populations 1 and 2. From equating administrations, estimates of the following for Population 1 are obtained: $\mu_1(X)$ = mean for X, $\sigma_1(X)$ = standard deviation for X, $\mu_1(V)$ = mean for V, $\sigma_1(V)$ = standard deviation for V, and $\sigma_1(X,V)$ = covariance between X and V; and for Population 2: $\mu_2(Y)$ = mean for Y, $\sigma_2(Y)$ = standard deviation for Y, $\mu_2(V)$ = mean for V, $\sigma_2(V)$ = standard deviation for V, and $\sigma_2(Y,V)$ = covariance between Y and V.

From the equating study, estimates of the following for Population 1 cannot be obtained: $\mu_1(Y)$ = mean for Y, $\sigma_1(Y)$ = standard deviation for Y, and $\sigma_1(Y,V)$ = covariance between Y and V; and for Population 2: $\mu_2(X)$ = mean for X, $\sigma_2(X)$ = standard deviation for X, and $\sigma_2(X,V)$ = covariance between X and V. This is so because Y is not administered to examinees from Population 1 and X is not administered to examinees from Population 2.

The assumptions used to arrive at expressions for these parameters distinguish the Tucker method from other linear methods for common item equating under the nonequivalent populations design. The Tucker method requires that the linear regression of X on V be identical for Populations 1 and 2. A similar assumption is required for the regression of Y on V. Let α represent a regression slope so that, for example,

$\alpha_1(X|V) = \sigma_1(X,V)/\sigma_1^2(V)$ is the slope for the linear regression of X on V for Population 1. Let β represent a regression intercept so that, for example, $\beta_1(X|V) = \mu_1(X) - \alpha_1(X|V)\mu_1(V)$ is the intercept for the linear regression of X on V for Population 1. The Tucker method requires that

$$\alpha_1(X|V) = \alpha_2(X|V) \quad , \tag{2}$$

$$\alpha_1(Y|V) = \alpha_2(Y|V) \quad , \tag{3}$$

$$\beta_1(X|V) = \beta_2(X|V) \quad , \tag{4}$$

and

$$\beta_1(Y|V) = \beta_2(Y|V) \quad . \tag{5}$$

In Tucker equating, it is also assumed that

$$\sigma_1^2(X)\,[1 - \rho_1^2(X,V)] = \sigma_2^2(X)\,[1 - \rho_2^2(X,V)] \quad , \tag{6}$$

and

$$\sigma_1^2(Y)\,[1 - \rho_1^2(Y,V)] = \sigma_2^2(Y)\,[1 - \rho_2^2(Y,V)] \quad , \tag{7}$$

where ρ^2 refers to a squared correlation. For example, $\rho_1^2(X,V) = \sigma_1^2(X,V)/[\sigma_1^2(X)\,\sigma_1^2(V)]$. This is sometimes referred to as the assumption that the variance errors of linearly estimating X from V as well as Y from V are the same for the two populations. Sometimes stronger assumptions are used for deriving these equations, such as those used by Braun and Holland (1982), but the assumptions listed in this paper are sufficient.

Given these assumptions, for Population 1

$$\mu_1(Y) = \mu_2(Y) + \alpha_2(Y|V)\,[\mu_1(V) - \mu_2(V)] \quad , \tag{8}$$

and

$$\sigma_1^2(Y) = \sigma_2^2(Y) + \alpha_2^2(Y|V)\,[\sigma_1^2(V) - \sigma_2^2(V)] \quad . \tag{9}$$

And, for Population 2

$$\mu_2(X) = \mu_1(X) - \alpha_1(X|V)\,[\mu_1(V) - \mu_2(V)] \quad , \tag{10}$$

and

$$\sigma_2^2(X) = \sigma_1^2(X) - \alpha_1^2(X|V)\,[\sigma_1^2(V) - \sigma_2^2(V)] \quad . \tag{11}$$

To arrive at the Tucker equating equation, expressions for the means and variances of X and Y for the synthetic population need to be obtained. These are expressible in terms of parameters for Populations 1 and 2 as follows:

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X) \quad , \tag{12}$$

$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y) \quad , \tag{13}$$

$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2 \quad , \tag{14}$$

and

$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2 \quad . \tag{15}$$

Substituting Equations 8 through 11 in Equations 12 through 15 gives

$$\mu_s(X) = \mu_1(X) - w_2\alpha_1(X|V)\,[\mu_1(V) - \mu_2(V)] \quad , \tag{16}$$

$$\mu_s(Y) = \mu_2(Y) + w_1\alpha_2(Y|V)\,[\mu_1(V) - \mu_2\,(V)] \quad , \tag{17}$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\alpha_1^2(X|V)\,[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\alpha_1^2(X|V)\,[\mu_1(V) - \mu_2(V)]^2 \quad , \tag{18}$$

and

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\alpha_2^2(Y|V)\,[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\alpha_2^2(Y|V)\,[\mu_1(V) - \mu_2(V)]^2 \quad , \tag{19}$$

where all parameters to the right of equal signs in Equations 16 through 19 are estimated directly using data from the study design. Equations 16 through 19 are entered into Equation 1 to produce the Tucker linear equating function.

It can be shown that Equations 16 through 19 produce counterparts of the Tucker method equation described by Angoff (1971, p. 580), if weights are chosen proportional to sample size—that is, $w_1 = n_1/(n_1 + n_2)$ and $w_2 = n_2/(n_1 + n_2)$, where n_1 and n_2 are the sample sizes of examinees included in the equating study from Populations 1 and 2, respectively. Gulliksen (1950, pp. 299–301) presented a version of the Tucker method that differs from Angoff's version. The present equations will result in counterparts of Gulliksen's by setting $w_1 = 1$ and $w_2 = 0$. To weight the populations equally, choose $w_1 = .5$ and $w_2 = .5$.

Large Sample Standard Errors

Kendall and Stuart (1977, pp. 246–247) presented a general method for approximating standard errors that is based on the Taylor expansion. This method is often referred to as the delta method. Lord (1950) presented standard errors of linear equating derived using the delta method under a variety of data collection designs, and many of these standard errors are reported by Angoff (1971). However, standard errors of Tucker equating are not presented in any of these sources.

In applying the delta method to standard errors of linear equating, Lord (1950) made what is referred to here as the *normality assumption*. For equating designs that require consideration of bivariate distributions, the normality assumption is that all of the central moments through order 4 of the score distributions are identical to those of a bivariate normal distribution, and for equating designs that require consideration of only univariate distributions, the normality assumption is that the central moments through order 4 of the score distributions are identical to those of a univariate normal distribution.

Recently, Braun and Holland (1982, pp. 32–35) derived standard errors using the delta method without making such a restrictive assumption for the situation in which randomly equivalent groups of examinees are administered the forms to be equated. Their resulting standard error expressions suggested that standard errors of equating based on the normality assumption may produce misleading results when score distributions are skewed or more peaked than a normal distribution. Because skewed distributions are typical in many testing programs, standard errors of Tucker equating are derived without the normality assumption in the present paper. Standard errors also are derived with the normality assumption for comparison purposes.

Let $\theta_1, \theta_2, \ldots, \theta_{10}$ be used as an alternate representation of $\mu_1(X)$, $\mu_1(V)$, $\sigma_1^2(X)$, $\sigma_1^2(V)$, $\sigma_1(X,V)$, $\mu_2(Y)$, $\mu_2(V)$, $\sigma_2^2(Y)$, $\sigma_2^2(V)$, and $\sigma_2(Y,V)$, respectively, and let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{10}$ represent their estimates. For example, $\hat{\theta}_1$ is an alternate representation of $\hat{\mu}_1(X)$. Let $\hat{\ell} \equiv \hat{\ell}(x)$ represent the estimated Tucker linear equating function arrived at by substituting estimates of parameters into Equations 16 through 19 and substituting these into Equation 1. Let ℓ_i' represent $\delta\hat{\ell}/\delta\hat{\theta}_i$ (the partial derivative of $\hat{\ell}$ with respect to $\hat{\theta}_i$) evaluated at $\theta_1, \theta_2, \ldots, \theta_{10}$. Then, by the delta method described by Kendall and Stuart (1977, pp. 246–247),

$$\mathrm{var}[\hat{\ell}(x)] \approx \sum_{i=1}^{10} \ell_i'^2\,\mathrm{var}(\hat{\theta}_i) + \sum_{i \neq j = 1}^{10}\sum \ell_i'\,\ell_j'\,\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j) \quad . \tag{20}$$

Because samples are independently drawn from Populations 1 and 2, the sampling covariances between each of the first five θ_is and each of the last five θ_is are zero. Thus,

$$\text{var}[\hat{\ell}(x)] \approx \sum_{i=1}^{10} \ell_i'^2 \, \text{var}(\hat{\theta}_i) + \sum_{i \neq j = 1}^{5} \sum \ell_i' \, \ell_j' \, \text{cov}(\hat{\theta}_i, \hat{\theta}_j) + \sum_{i \neq j = 5}^{10} \sum \ell_i' \, \ell_j' \, \text{cov}(\hat{\theta}_i, \hat{\theta}_j) \quad . \tag{21}$$

The partial derivatives (ℓ_i's) necessary for Equation 21 are shown in Table 1. For this table, $z_x = [x - \mu_s(X)]/\sigma_s(X)$. All other notation has been defined previously. The sampling variances and covariances for Equation 21 can be obtained from Table 2. In this table, n refers to sample size. (Note that the variables X and V in Table 2 are general.) By substituting the partial derivatives from Table 1 and the sampling variances and covariances from the "general" column in Table 2 into Equation 21, the equation for the variance error of Tucker equating results. (Use the "normal" column of Table 2 for the variance error under the normality assumption.) Note that the standard error for Tucker equating is $\text{SE}[\hat{\ell}(x)] = \{\text{var}[\hat{\ell}(x)]\}^{1/2}$.

As an example of how to proceed, refer to the first term in the first summation in Equation 21, which is $\ell_1'^2 \, \text{var}(\hat{\theta}_1)$. From Table 1, $\ell_1'^2$ is $\sigma_s^2(Y)/\sigma_s^2(X)$, and from Table 2, $\text{var}(\hat{\theta}_1) \equiv \text{var}[\hat{\mu}_1(X)] = \sigma_1^2(X)/n_1$. Note that this term is the same under the general or the normality assumption. As another example, refer to the second term in the second summation of Equation 21, which is $\ell_1'\ell_3' \, \text{cov}(\hat{\theta}_1, \hat{\theta}_3)$. From Table 1, $\ell_1' \, \ell_3' = [-\sigma_s(Y)/\sigma_s(X)]\{-z_x\sigma_s(Y)/[2\sigma_s^2(X)]\}$, and from Table 2, $\text{cov}(\hat{\theta}_1,\hat{\theta}_3) \equiv \text{cov}[\hat{\mu}_1(X), \hat{\sigma}_1^2(X)] \approx E[X-\mu_1(X)]^3/n_1$ under general conditions. From the table, this term would be zero under the normality assumption.

The standard errors are not written here in more detail than Equation 21 because their full form is too cumbersome. However, the standard errors are easily programmed using a computer.

Clearly, the standard error expression is complicated. For this reason, it is difficult to make general interpretive statements. One such observation, however, is that if the sample sizes for the two groups are equal, then there is a simple relationship between the variance error and sample size—namely, the magnitude of the variance error is inversely proportional to sample size. For example, a doubling of the sample size will lead to a halving of the variance error.

When standard errors of Tucker equating are estimated, parameter estimates must be used to calculate the derivatives shown in Table 1 and the sampling variances and covariances shown in Table 2. Under the normality assumption, means, variances, and covariances need to be estimated to obtain the sampling variances and covariances in Table 2. Under nonnormality, skewness, kurtosis, and several higher order cross-product moments also need to be estimated.

Lord (1955a) developed a linear method for equating under a common item design in which the examinees are assumed to be random samples from a single population. To derive the estimates for this method, Lord also assumed that test scores are normally distributed. Angoff (1971, p. 580) indicated that the estimation equations for Lord's method are the same as those for the Tucker method. In the notation of the present paper, the estimation equations would be the same for the two methods if $n_1 = n_2$ and $w_1 = w_2 = .5$. However, the standard errors for Lord's method, as presented in Angoff (1971, p. 577) and Lord (1950), are not, in general, the same as those for the Tucker method. To obtain the standard errors of Lord's method from the results in the present paper, Equation 21 is used in conjunction with the "normal" column in Table 2 and $n_1 = n_2$. Then, Table 1 is entered with $w_1 = w_2$, $\mu_1(V) = \mu_2(V)$, and $\sigma_1(V) = \sigma_2(V)$. The "normal" column in Table 2 is used because Lord (1955a) assumed normality in his derivation of the method. The restrictions $\mu_1(V) = \mu_2(V)$ and $\sigma_1(V) = \sigma_2(V)$ are needed because Populations 1 and 2 are assumed to be the same population in Lord's method.

To compare the standard errors for the two methods, note that the partial derivatives for Lord's (1950) method can be arrived at by replacing in Table 1 all occurrences of $\mu_1(V) - \mu_2(V)$ by zero, and all occurrences of $\sigma_s^2(V)/\sigma_1^2(V)$ and $\sigma_s^2(V)/\sigma_2^2(V)$ by 1. This greatly simplifies the partial derivatives and

Table 1
Partial Derivatives of Tucker Linear Equating Equation
With Respect to Each Sample Statistic and Evaluated at the Parameters

Statistic	Derivatives Evaluated at Parameters
$\hat{\mu}_1(X)$	$-\sigma_s(Y)/\sigma_s(X)$
$\hat{\mu}_1(V)$	$w_2\sigma_s(Y)\alpha_1(X\|V)/\sigma_s(X) + w_1w_2z_x\,\alpha_2^2(Y\|V)[\mu_1(V) - \mu_2(V)]/\sigma_s(Y)$ $-w_1w_2\sigma_s(Y)z_x\,\alpha_1^2(X\|V)[\mu_1(V) - \mu_2(V)]/\sigma_s^2(X) + w_1\alpha_2(Y\|V)$
$\hat{\sigma}_1^2(X)$	$-z_x\sigma_s(Y)/[2\sigma_s^2(X)]$
$\hat{\sigma}_1^2(V)$	$-w_2\sigma_s(Y)\,\alpha_1(X\|V)[\mu_1(V) - \mu_2(V)]/[\sigma_s(X)\sigma_1^2(V)]$ $+w_1z_x\alpha_2^2(Y\|V)/[2\sigma_s(Y)]$ $-\sigma_s(Y)z_x\alpha_1^2(X\|V)[1 + w_1 - 2\sigma_s^2(V)/\sigma_1^2(V)]/[2\sigma_s^2(X)]$
$\hat{\sigma}_1(X,V)$	$w_2\sigma_s(Y)[\mu_1(V) - \mu_2(V)]/[\sigma_s(X)\sigma_1^2(V)]$ $-\sigma_s(Y)z_x\,\alpha_1(X\|V)[\sigma_s^2(V)/\sigma_1^2(V) - 1]/\sigma_s^2(X)$
$\hat{\mu}_2(Y)$	1
$\hat{\mu}_2(V)$	$-w_2\sigma_s(Y)\,\alpha_1(X\|V)/\sigma_s(X) - w_1w_2z_x\,\alpha_2^2(Y\|V)[\mu_1(V) - \mu_2(V)]/\sigma_s(Y)$ $+w_1w_2\sigma_s(Y)z_x\,\alpha_1^2(X\|V)[\mu_1(V) - \mu_2(V)]/\sigma_s^2(X) - w_1\alpha_2(Y\|V)$
$\hat{\sigma}_2^2(Y)$	$z_x/[2\sigma_s(Y)]$
$\hat{\sigma}_2^2(V)$	$-w_2\sigma_s(Y)z_x\,\alpha_1^2(X\|V)/[2\sigma_s^2(X)] - w_1\alpha_2(Y\|V)[\mu_1(V) - \mu_2(V)]/\sigma_2^2(V)$ $+z_x\alpha_2^2(Y\|V)[1 + w_2 - 2\sigma_s^2(V)/\sigma_2^2(V)]/[2\sigma_s(Y)]$
$\hat{\sigma}_2(Y,V)$	$z_x\alpha_2(Y\|V)[\sigma_s^2(V)/\sigma_2^2(V) - 1]/\sigma_s(Y)$ $+w_1[\mu_1(V) - \mu_2(V)]/\sigma_2^2(V)$

suggests that the standard errors for the Tucker method are quite different from those for Lord's method, though the relationship between the standard errors appears to be complicated. However, when Populations 1 and 2 are more similar, it does seem that the standard errors for the two methods also will be more similar.

Table 2
Sampling Variances and Covariances of Bivariate Moments

Statistic(s)	Sampling Variance or Covariance--General	Sampling Variance or Covariance-Normal Distribution
$\mathrm{var}[\hat{\mu}(X)]$	$\sigma^2(X)/n$	$\sigma^2(X)/n$
$\mathrm{var}[\hat{\sigma}^2(X)]$	$\{E[X - \mu(X)]^4 - \sigma^4(X)\}/n$	$2\sigma^4(X)/n$
$\mathrm{var}[\hat{\sigma}(X,V)]$	$\{E[X - \mu(X)]^2[V - \mu(V)]^2 - \sigma^2(X,V)\}/n$	$[\sigma^2(X)\sigma^2(V) + \sigma^2(X,V)]/n$
$\mathrm{cov}[\hat{\mu}(X) , \hat{\mu}(V)]$	$\sigma(X,V)/n$	$\sigma(X,V)/n$
$\mathrm{cov}[\hat{\mu}(X) , \hat{\sigma}^2(X)]$	$E[X - \mu(X)]^3/n$	0
$\mathrm{cov}[\hat{\mu}(X) , \hat{\sigma}^2(V)]$	$E[X - \mu(X)][V - \mu(V)]^2/n$	0
$\mathrm{cov}[\hat{\mu}(X) , \hat{\sigma}(X,V)]$	$E[X - \mu(X)]^2[V - \mu(V)]/n$	0
$\mathrm{cov}[\hat{\sigma}^2(X) , \hat{\sigma}^2(V)]$	$\{E[X - \mu(X)]^2[V - \mu(V)]^2 - \sigma^2(X)\sigma^2(V)\}/n$	$2\sigma^2(X,V)/n$
$\mathrm{cov}[\hat{\sigma}^2(X) , \hat{\sigma}(X,V)]$	$\{E[X - \mu(X)]^3 [V - \mu(V)] - \sigma^2(X)\sigma(X,V)\}/n$	$2\sigma(X,V)\sigma^2(X)/n$

Note: The terms in the body of the table were adapted from Kendall and Stuart (1977, pp. 85, 245, 250) and are typically based on large sample theory. E refers to expected value.

Computer Simulation

Method

A computer simulation was conducted to study the behavior of the estimated standard errors. Score distributions were simulated to reflect the score distributions of test forms from two different testing programs. The distributions for two test forms model those in a particular professional certification program. (Real data for real forms of a test in this program are used in a subsequent illustration.) These distributions are negatively skewed, and the simulation based on these distributions is referred to as the *nonsymmetric simulation*. Distributions for two forms of a second test are modeled after the mean, standard deviation, skewness, and kurtosis found in an achievement testing program. The simulation based on these distributions is referred to as the *nearly symmetric simulation*. The distributions in the nearly symmetric simulation were flatter than a normal distribution. (Lord, 1955b, surveyed distributions for a variety of tests and found that symmetric test score distributions tend to be flatter than the normal distribution, and he referenced theoretical discussions of this issue.) Sample size for the simulations was 100 examinees per form and 250 examinees per form. These sample sizes were chosen because they are the smallest that are typically used for equating; usually sample size is substantially larger than these. For purposes of the simulation, the beta-binomial model, which has been shown to fit a variety of test score distributions quite well, was used (see Lord & Novick, 1968, chap. 23).

For the nonsymmetric simulation, the beta true-score distributions of X and V for Population 1 were assigned parameters 10.5 and 3.0. For Population 2, the beta true-score distributions of Y and V were assigned parameters 9.5 and 3.0. The numbers of items in these simulated test forms were 125 for X and Y and 30 for V. For the nearly symmetric simulation, the beta true-score distributions of X and V for Population 1 were assigned parameters 6.0 and 6.2. For Y and V for Population 2, the parameters assigned were 5.4 and 5.2. The numbers of items in these simulated tests were 52 for X and Y and 15 for V.

Population means, standard deviations, skewness indices, and kurtosis indices of observed scores are shown in Table 3 for the simulated test forms. The nonsymmetric distributions have means near 75% of the items answered correctly, are negatively skewed, and are more peaked than a normal distribution. The nearly symmetric distributions have means near 50% of the items answered correctly, are nearly symmetric, and are less peaked than a normal distribution.

For the simulation, let k_x represent the number of items on X, k_y the number of items on Y, and k_v the number of items on V. Also, for the simulation $k_x = k_y$. Define $k_g = k_x - k_v$. Because an internal set of common items was being simulated, k_g represents the number of items in X and Y that are not common, and k_v represents the number of common items. Note that the standard error equations hold whether V is internal or external to X and Y.

The following steps were used to simulate pairs of X and V scores:

1. Randomly generate a beta variate from parameters associated with Population 1. This beta variate represents a proportion-correct true score. (International Mathematical and Statistical Libraries, IMSL, 1982, subroutine GGBTR was used to generate the beta variate.)
2. Given the true score, p, in step 1, randomly generate a binomial variate for k_v trials. This variate represents observed score on V. (IMSL, 1982, subroutine GGBN was used to generate binomial variates.)
3. Randomly generate a binomial variate with parameter p based on k_g trials. This variate represents observed score on the noncommon items.
4. Add together the binomial variates from steps 2 and 3. This sum represents observed score on the total test form, X.
5. Repeat steps 1 through 4 n times, where n represents the sample size used in the simulation. This results in a set of n pairs of observed scores for X and V.

Table 3
Population Means, Standard Deviations, Skewness, and Kurtosis for Simulated Observed Score Distributions

Variable	Population	Number of Items	Mean	Standard Deviation	Skewness	Kurtosis
			Nonsymmetric			
X	1	125	97.22	14.37	-0.66	3.24
Y	2	125	93.75	15.72	-0.60	3.09
V	1	30	23.33	3.94	-0.67	3.23
V	2	30	22.50	4.26	-0.60	3.09
			Nearly Symmetric			
X	1	52	25.57	7.95	0.02	2.60
Y	2	52	26.49	8.37	-0.02	2.55
V	1	15	7.39	2.78	0.02	2.55
V	2	15	7.64	2.88	-0.02	2.51

Note: Skewness is Pearson's $\sqrt{\beta_1}$ and kurtosis is Pearson's β_2 index.

Next, by substituting Y for X and Population 2 for Population 1 in the above steps, n pairs of observed scores were generated for Population 2 using the appropriate beta parameters. At this point, there were n pairs of scores on X and V for Population 1 and n pairs of scores on Y and V for Population 2. Based on these simulated data, a Form Y equivalent of each Form X integer score was obtained using Tucker equating with $w_1 = w_2 = .5$. Also, standard errors of equating were estimated for each X (integer) score based on the delta method with the normality assumption as well as the delta method without the normality assumption. This process was replicated 500 times. The "true" standard error of equating for a given integer score on X was defined here as the standard deviation of Form Y equivalents of that score over the 500 replications. The nonnormal delta method standard error associated with each X (integer) score is the mean delta method standard error derived without the normality assumption over the 500 replications. The normal delta method standard error was defined similarly.

Results

Nonsymmetric and symmetric simulations were each conducted using sample sizes of 100 and 250 simulated examinees per form. The "true," nonnormal, and normal standard errors at selected score points are shown in Table 4. For example, the top row gives standard errors of Form Y equivalents of a Form X score of 120 based on the nonsymmetric simulation with a sample size of 250 examinees per test form. In this row, the "true" standard error is 1.01, the nonnormal standard error is .96, and the normal standard error is 1.12. Root mean squared errors (RMSE) in estimating the standard errors are shown also.

Table 4 illustrates the dependence of the standard errors on score; the standard errors are smaller near the mean than at the extremes. Table 4 also illustrates that the standard errors are smaller for larger sample sizes.

Table 4
Standard Errors of Tucker Equating for Two Simulated Tests and at Two Sample Sizes

Score on Form X	Standard Error			RMSE in Estimating Standard Error	
	"True"	Nonnormal	Normal	Nonnormal	Normal
		Nonsymmetric $n_1=n_2=250$			
120	1.01	0.96	1.12	.08	.13
110	0.70	0.68	0.81	.04	.12
100	0.58	0.59	0.63	.02	.06
90	0.75	0.78	0.69	.05	.07
80	1.09	1.10	0.94	.08	.15
70	1.48	1.47	1.28	.11	.21
60	1.89	1.87	1.65	.15	.26
50	2.32	2.27	2.03	.19	.31
		Nonsymmetric $n_1=n_2=100$			
120	1.55	1.49	1.76	.16	.26
110	1.07	1.06	1.27	.09	.22
100	0.94	0.93	0.99	.06	.08
90	1.28	1.21	1.08	.14	.22
80	1.85	1.71	1.48	.25	.39
70	2.49	2.28	2.00	.36	.52
60	3.16	2.89	2.58	.47	.63
50	3.84	3.51	3.19	.57	.72
		Nearly Symmetric $n_1=n_2=250$			
50	1.12	1.12	1.20	.07	.10
40	0.73	0.74	0.78	.05	.06
30	0.44	0.45	0.45	.02	.02
20	0.46	0.46	0.48	.02	.03
10	0.78	0.77	0.82	.05	.06
0	1.16	1.15	1.25	.07	.11
		Nearly Symmetric $n_1=n_2=100$			
50	1.82	1.77	1.89	.20	.17
40	1.20	1.16	1.23	.13	.11
30	0.73	0.70	0.71	.06	.05
20	0.77	0.73	0.75	.07	.06
10	1.27	1.22	1.31	.13	.11
0	1.90	1.83	1.98	.20	.18

In addition, Table 4 provides evidence of the accuracy of the standard errors for the nonsymmetric and nearly symmetric simulations. The closer the nonnormal and normal standard errors are to the "true" standard errors, the more accurate they are considered to be. By this criterion, the nonnormal standard errors are more accurate than the normal standard errors for the nonsymmetric simulation at both sample sizes. For the nearly symmetric simulation, the nonnormal standard errors are more accurate than the normal standard errors at a sample size of 250. In addition, for both simulations the nonnormal standard errors are more accurate for the larger than for the smaller sample sizes. For the normal standard errors, this desirable property does not appear to hold in the nearly symmetric simulation.

RMSE in estimating the delta method standard errors also are shown in Table 4. To calculate RMSE, the variance of the estimated standard errors over the 500 replications was computed and added to the squared difference between the "true" standard error and the delta method standard error. The square root of this sum is the RMSE. The RMSE is a measure of the variability in estimating standard errors. Recall that the estimation of the normal standard errors requires estimation of only means, variances, and covariances, whereas the estimation of the nonnormal standard errors requires the estimation of these parameters as well as higher-order central moments and cross-product moments. Because higher-order moments and cross-product moments may be difficult to estimate precisely due to sampling variability, the nonnormal standard errors may be more difficult to estimate than the normal ones. However, for all but the nearly symmetric simulation with sample size of 100 in Table 4, the RMSE is smaller for the nonnormal standard errors than for the normal standard errors.

In summary, the results of the simulation indicate that the nonnormal standard errors are more accurate than those based on the normality assumption. The simulation also suggests that as sample size increases, the nonnormal standard errors become closer to the "true" standard errors. Thus, as sample size increases beyond 250 examinees per form, the accuracy of the nonnormal standard errors is expected to increase further. This property does not appear to hold for the standard errors derived using the normality assumption.

Bootstrap Standard Errors

Even though the simulation provides evidence of the behavior of the standard errors, a study of the delta method standard errors of equating using actual test data seemed desirable. Efron (1982) described an alternative to the delta method which he referred to as the bootstrap, and he presented a variety of examples in which the bootstrap resulted in standard errors that were more accurate for small sample situations than those based on the delta method. In the present paper, the bootstrap is used as a confirmatory procedure for evaluating the accuracy of the nonnormal and normal standard errors.

Method

The computation of bootstrap standard errors requires extensive resampling from the sample data. Thus, a high-speed computer is essential. In general, to compute bootstrap standard errors, a random sample is drawn with replacement from the sample data at hand, the statistic of interest is calculated, and this process is repeated a large number of times. The bootstrap standard error is the standard deviation of the computed values of the statistic over repetitions of the process. The following steps are used to estimate bootstrap standard errors of Tucker equating.

1. Begin with the n_1 examinees from Population 1 with scores on X and V and the n_2 examinees from Population 2 with scores on Y and V.
2. Draw a random sample with replacement of size n_1 from the sample of n_1 examinees from Population 1.

The sampling involves drawing pairs of X and V scores. Because the sampling is with replacement, a particular examinee's score pair could be easily chosen more than once.

3. Draw a random sample with replacement of size n_2 from the sample data of n_2 examinees from Population 2.
4. Estimate the Tucker equivalent at x using the data from the random samples drawn in steps 2 and 3, and refer to this estimate as $\hat{\ell}_b(x)$.
5. Repeat steps 2 through 4 B times obtaining bootstrap estimates $\hat{\ell}_1(x), \hat{\ell}_2(x), \ldots, \hat{\ell}_B(x)$. Approximate the standard error by:

$$SE_{Boot}[\hat{\ell}(x)] = \{\sum_{b=1}^{B} [\hat{\ell}_b(x) - \hat{\ell}_\cdot(x)]^2/(B-1)\}^{1/2} \quad , \tag{22}$$

where

$$\hat{\ell}_\cdot(x) = \sum_{b=1}^{B} \hat{\ell}_b(x)/B \quad . \tag{23}$$

These procedures can be applied at any x.

Real Data Example

Data from Forms X and Y of a 125-item multiple-choice professional certification testing program were used in this example. Form X was administered to 773 examinees from Population 1 and Form Y to 795 examinees from Population 2; the forms were administered one year apart. The two forms contained a common set of 30 items, referred to as V. The summary statistics shown in Table 5 indicate that the average examinee correctly answered approximately 77% of the items, the score distributions are markedly skewed, and the distributions are more peaked than a normal distribution.

Results from Tucker equating with $w_1 = w_2 = .5$ and standard errors of equating are shown in Table 6. Consider a Form X raw score of 100 in the first column of the table. This score has a percentile rank of 54.7 and a Form Y equivalent of 102.7. The standard error of this equivalent is .33 under normality assumptions, .29 without these assumptions, and .28 using the bootstrap. A ± one standard error band for the Form Y equivalent of a Form X score of 100 is 102.7 ± .29 or approximately (102.4, 103.0) for the standard errors derived without the normality assumption.

Table 5
Raw Score Summary Statistics for Forms X and Y and Common Items V for a Professional Certification Examination

Variable	Group	Mean	Standard Deviation	Skewness	Kurtosis
X	1	95.75	13.38	-1.03	3.91
Y	2	96.84	13.37	-1.00	3.89
V	1	23.18	4.05	-0.84	3.48
V	2	22.54	4.31	-0.79	3.47

Note: Skewness is Pearson's $\sqrt{\beta_1}$ and kurtosis is Pearson's β_2 index. Sample sizes are 773 and 795 for Groups 1 and 2, respectively. There are 125 items on X and Y and 30 items on V.

Table 6
Standard Errors of Tucker Equating
for a Professional Certification Program

Form X Raw Score	Percentile Rank In Group 1	Form Y Equivalent	Standard Errors		
			Normality	Nonnormality	Bootstrap[a]
125	100.0	126.5	0.71	0.67	0.69
120	99.9	121.7	0.61	0.56	0.58
115	98.0	116.9	0.53	0.47	0.48
110	90.1	112.2	0.44	0.38	0.39
105	73.4	107.4	0.38	0.32	0.32
100	54.7	102.7	0.33	0.29	0.28
95	40.2	97.9	0.31	0.30	0.30
90	27.3	93.1	0.32	0.36	0.35
85	18.6	88.4	0.37	0.44	0.44
80	12.1	83.6	0.44	0.54	0.53
75	8.9	78.9	0.52	0.64	0.64
70	6.0	74.1	0.61	0.74	0.75
65	3.8	69.4	0.70	0.85	0.86
60	2.3	64.6	0.80	0.96	0.97
55	0.6	59.8	0.90	1.08	1.08
50	0.0	55.1	1.00	1.19	1.20

[a]Based on B=1000 bootstrap replications.

In general, the standard errors are smallest near the mean and become larger farther away from the mean. The standard errors under the normality assumption are slightly larger at the higher scores and are smaller at the lower scores than those derived without the normality assumption and those calculated by the bootstrap. Standard errors for the bootstrap and the delta method without the normality assumption are nearly identical.

For this testing program the passing score is usually close to a raw score of 80, so equating error is crucial in this region. From Table 6, at a raw score of 80, the delta method standard error of equating is .44 under the normality assumption and .54 without such an assumption. The error variances are, respectively, .19 ($.44^2$) and .29 ($.54^2$). To place the difference between these in perspective, consider that at a score of 80, the error variance under normality is only 66% [100(.19)/.29] of the size of the error variance under the less restrictive assumption. Based on the less restrictive assumption, these results suggest that instead of needing approximately 780 examinees, approximately 1,182 (780/.66) examinees are needed to obtain the precision implied by the error variance based on the normality assumption, which is a substantial difference. The close agreement between the bootstrap standard errors and the delta method standard errors derived without the normality assumption, in combination with the findings from the previously discussed nonnormal simulations shown in the RMSE column in Table 4, suggest that, for this real data example, the sample size estimates using the standard errors based on the normality assumption are not likely as accurate as those based on the less restrictive assumption.

Discussion

The results of the computer simulation indicate that for Tucker equating the standard errors derived without the normality assumption are more accurate than those derived with the normality assumption,

for sample sizes of 250 or more examinees per test form. The results also indicate that the standard errors derived with the normality assumption may be acceptable when test score distributions are nearly symmetric, but these standard errors appear to be inadequate for nonsymmetric distributions. The results of the real data example indicate that the standard errors with the normality assumption may suggest substantially more equating precision in the crucial range than is actually the case.

In the real data example, the bootstrap standard errors are very similar to the delta method standard errors derived without the normality assumption, which are preferable to the bootstrap standard errors for reasons of cost and ease of computation. Still, the results are encouraging for the use of the bootstrap in equating contexts. Ultimately, the bootstrap may prove useful for estimating standard errors of equating in complicated situations such as in chains of dependent equipercentile equatings or in smoothed equipercentile equating.

The standard errors derived in this paper index equating error that is due to examinee sampling. Error that results from a failure to meet the assumptions required for Tucker equating is not reflected in the standard errors. Braun and Holland (1982) suggested some procedures for checking these assumptions, though they indicated that not all of the assumptions are testable without collecting additional data. Further research should be conducted on the effects of the violations of assumptions. One such study might involve equating a test to itself and using a resampling plan to separate error due to examinee sampling from error due to violation of the statistical assumptions.

The assumptions required in Tucker equating seem most reasonable for testing programs in which: (1) examinee populations do not change much from test date to test date; (2) the content balance of the set of common items is very similar to the content balance of the total test forms, and the total test forms are each built to the same specifications; and (3) the statistical characteristics of the set of common items are very similar to the statistical characteristics of the total test forms, and the total test forms are similar to one another in statistical characteristics. Characteristics 2 and 3 are most readily achieved for testing programs in which tests are constructed from a large pool of items with item statistics that are accurately estimated from pretesting or previous use.

For equating under the common item nonequivalent populations design, other methods have been suggested for situations in which the Tucker method assumptions are not likely to hold. Angoff (1971, p. 582) suggested that the Levine equally reliable method be used when groups differ widely in ability. The Levine method, however, requires the strong assumption that the correlations between true scores on X and Y, X and V, and Y and V are all 1.0. This assumption seems reasonable only if the test forms are very similar. Angoff (p. 581) described an equipercentile method that is appropriate when the equating relationship is curvilinear and the examinee groups are similar in ability. Jarjoura and Kolen (in press) and Kolen and Jarjoura (1985) indicated that this method is promising. Item response theory methods also are available. These methods require that the tests be unidimensional; such an assumption may be problematic for heterogeneous educational tests. In addition, fully acceptable item response theory methods for equating observed scores have yet to be developed (Lord, 1982, pp. 147–148).

Conclusions

Standard errors of Tucker equating were derived without the normality assumption because many testing programs produce score distributions that deviate markedly from a normal distribution. For example, professional certification testing programs often produce markedly negatively skewed score distributions that result, in part, because the mean score on such examinations is often in the range of 65% to 80% of the items correct. Many of the testing programs that produce skewed distributions are equated using linear methods under the common item nonequivalent populations design with sample sizes of 250 or more examinees per test form. The results of the analyses in this paper indicate that the standard errors of Tucker equating derived without the normality assumption are sufficiently accurate for such programs.

References

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

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.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia PA: Society for Industrial and Applied Mathematics.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

International Mathematical and Statistical Libraries. (1982). *Reference manual* (9th ed.). Houston: Author.

Jarjoura, D., & Kolen, M. J. (in press). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics*.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). New York: Macmillan.

Kolen, M. J., & Jarjoura, D. (1985). *Analytic smoothing for equipercentile equating under the common item nonequivalent populations design* (ACT Technical Bulletin No. 47). Iowa City IA: The American College Testing Program.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (RB-50-48). Princeton NJ: Educational Testing Service.

Lord, F. M. (1955a). Equating test scores—A maximum likelihood solution. *Psychometrika, 20,* 193–200.

Lord, F. M. (1955b). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15,* 383–389.

Lord, F. M. (1982). Item response theory equating—A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 141–148). New York: Academic Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Acknowledgments

The author thanks David Jarjoura, Robert L. Brennan, and Ronald T. Cope for their comments and suggestions.

Author's Address

Send requests for reprints and further information to Michael J. Kolen, The American College Testing Program, P.O. Box 168, Iowa City IA 52243, U.S.A.