Standard Errors of A Chain of Linear Equatings

Lingjia Zeng, Bradley A. Hanson, and Michael J. Kolen
American College Testing

A general delta method is described for computing the standard error (SE) of a chain of linear equatings. The general delta method derives the SEs directly from the moments of the score distributions obtained in the equating chain. The partial derivatives of the chain equating function needed for computing the SEs are derived numerically. The method can be applied to equatings using the common-items nonequivalent populations design. Computer simulations were conducted to evaluate the SEs of a chain of two equatings using the Levine and Tucker methods. The general delta method was more accurate than a method that assumes the equating processes in the chain are statistically independent. *Index terms: chain equating, delta method, equating, linear equating, standard error of equating.*

In many testing programs, test forms are different from one administration to another for security reasons. However, there is often a need to compare scores obtained from taking different forms of a test. Equating is used to ensure that scores earned on alternate forms are comparable. In a commonly used equating scheme, newer forms are equated back to the base form through a chain of equatings. Number-correct scores from a particular form thus can be transformed to the scale of any other form by performing a chain of equatings. Because equating is based on the data obtained by administering test forms to samples drawn from populations of potential examinees, the results of equating are subject to sampling errors. The standard error (SE) is a useful statistic for quantifying the sampling errors that propagate in a chain of equatings.

SEs of equating (SEEq) for different equating designs have been reported by a number of authors (e.g., Angoff, 1984; Braun & Holland, 1982; Hanson, Zeng, & Kolen, 1993; Kolen, 1985; Lord, 1950). However, only a few studies have been conducted on the SEs of a chain of equatings. Braun & Holland (1982) provided a formula for computing the SEs of a chain of equatings; they assumed that each equating conducted in the chain was statistically independent of every other equating. In many testing programs, especially those using the common-items nonequivalent populations design (CINPD), the equatings conducted in a chain are not statistically independent. Under the CINPD, two groups of examinees sampled from different populations are each administered different test forms that have a subset of items in common. In such situations, procedures other than those described by Braun and Holland must be used.

A general method for computing SEs of a chain of linear equatings is introduced here. In the proposed method, the SEs are derived directly from the moments of score distributions obtained in the chain of equatings; therefore, the assumption of statistical independence of the equatings is not required. Thus, the proposed method—the general delta method—can be applied to equating using the CINPD.

An Equating Chain

Consider a chain of equating g test forms, which are designated Form 1, Form 2, ..., Form g. Assume that Form g is equated to Form $g-1$, which is equated to Form $g-2$, and so on, so that through a *chain of equatings*, Form g is eventually equated to Form 1. Let $l_{j,k}(x_k)$ be a linear function that converts score x_k on

0146-6216/94/040369-10$1.75

Form k to the scale of Form j, where $j \neq k$. For example, $l_{1,2}(x_2)$ is used to convert scores on Form 2 to the scale for Form 1. Using this notation, a chain of equatings can be expressed as a composite function composed of two or more nested linear functions as follows:

$$l_{1,3}(x_3) = l_{1,2}\left[l_{2,3}(x_3)\right] ,$$

$$l_{1,4}(x_4) = l_{1,2}\left\{l_{2,3}\left[l_{3,4}(x_4)\right]\right\} ,$$

$$\vdots$$

$$l_{1,g}(x_g) = l_{1,2}\left\{l_{2,3}\left[\ldots l_{g\text{-}1,g}(x_g)\ldots\right]\right\} . \tag{1}$$

The outcome of a chain of equatings is a single linear equation that is used to convert Form g scores to the scale of Form 1.

The relationship between two adjacent equatings in a chain is determined by the data collection design used in the equatings. Suppose in a chain of equating g forms, each pair of two adjacent forms is equated using the data obtained by administering the two forms to two randomly equivalent samples of examinees drawn from the same population. In this situation, because samples are independent of each other from one equating to the next, the equatings in the chain are statistically independent of each other. Chains of statistically independent equatings are common in practice. For example, the forms of the ACT Assessment Program (American College Testing Program, 1988) are equated through an independent chain of equatings using a random groups design.

Suppose a chain of g equatings is conducted using the CINPD. Each form is administered to only one sample of examinees from a population. Data obtained from one sample of examinees may be used in two equatings. For example, data obtained from the same sample is used in equating Form 2 to Form 1 and in equating Form 3 to Form 2. The moments of the score distribution for Form 2 are used in the two equatings. As a result, the two adjacent equatings in the chain are no longer statistically independent. That is, the equatings in a chain are statistically dependent when the CINPD is used.

Standard Errors of a Chain of Equatings

A Method Assuming Independent Equatings

The SEEq is defined as the square root of the sampling variance of equating. Suppose $\hat{l}_{1,3}(x_3)$ is the estimated composite function of a chain of equatings that converts score x_3 of Form 3 to the scale of Form 1. If $\hat{l}_{1,3}(x_3)$ is composed of two statistically independent functions, $\hat{l}_{1,2}(x_2)$ and $\hat{l}_{2,3}(x_3)$, then the sampling variance of $\hat{l}_{1,3}(x_3)$ can be estimated by the following formula described by Braun & Holland (1982):

$$\text{var}\left[\hat{l}_{1,3}(x_3)\right] = \text{var}\left\{\hat{l}_{1,2}\left[\hat{l}_{2,3}(x_3)\right]\right\} + \left\{l'_{1,2}\left[\hat{l}_{2,3}(x_3)\right]\right\}^2 \text{var}\left[\hat{l}_{2,3}(x_3)\right] , \tag{2}$$

where

$\text{var}\left\{\hat{l}_{1,2}\left[\hat{l}_{2,3}(x_3)\right]\right\}$ is the sampling variance of the equating function $\hat{l}_{1,2}$ at score level $\hat{l}_{2,3}(x_3)$,

$\text{var}\left[\hat{l}_{2,3}(x_3)\right]$ is the sampling variance of the equating function $\hat{l}_{2,3}$ at score level x_3, and

$l'_{1,2}\left[\hat{l}_{2,3}(x_3)\right]$ is the derivative of $l_{1,2}$ evaluated at the value $\hat{l}_{2,3}(x_3)$.

Because the derivative of a linear equating function is generally close to 1.0, Equation 2 suggests that the variance of a chain of equatings is approximately the sum of the variances of all its individual equatings provided all the equatings conducted in the chain are statistically independent.

The General Delta Method

If the assumption of independence cannot be met, such as in a testing program that uses the CINPD, the delta method described by Kendall & Stuart (1977) may be used. The delta method is a general numerical approach derived from Taylor's series that estimates the sampling variance for a function of a number of random variables. Suppose that the chain equating function $\hat{l}_{1,g}(x_g)$ contains s random variables (the estimates of sample moments, $\hat{\theta}_i$, $i = 1, ..., s$). According to Kendall and Stuart (pp. 246–247), the sampling variance of $\hat{l}_{1,g}(x_g)$ can be approximated as:

$$\mathrm{var}\left[\hat{l}_{1,g}(x_g)\right] \approx \sum_{i=1}^{s}\left(\frac{\partial l_{1,g}}{\partial\theta_i}\right)^2 \mathrm{var}(\hat{\theta}_i) + \sum_{i=1}^{s}\sum_{j\neq i=1}^{s}\frac{\partial l_{1,g}}{\partial\theta_i}\frac{\partial l_{1,g}}{\partial\theta_j}\mathrm{cov}(\hat{\theta}_i,\hat{\theta}_j), \qquad (3)$$

where

$\partial l_{1,g}/\partial\theta_i$ is the partial derivative of the composite function $l_{1,g}$ with respect to variable θ_i,

$\mathrm{var}(\theta_i)$ is the variance of $\hat{\theta}_i$, and

$\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j)$ is the covariance of moments $\hat{\theta}_i$ and $\hat{\theta}_j$.

Because a number of independent samples are used in $\hat{l}_{1,g}$, the sampling covariance of two moments that are not from the same sample is 0.0. Thus, some of the terms in Equation 3 that involve moments from two samples are zero terms and can be eliminated using reformulation.

Consider a typical form, Form k. Let the random variable X_k refer to a score on Form k (for an internal link, X_k refers to a score on all items; for an external link, only noncommon items). (For an internal link, scores on common items contribute to the total scores for both forms; for an external link, scores on common items do not contribute to the total score.) Assume that Form k is involved in an equating chain in which the CINPD is used. In this case, for $2 \le k \le g$, Form k has a set of items that are in common with Form $k-1$. For examinees who were administered Form k ($2 \le k \le g$), let the random variable V_k refer to the score on the items that are in common with Form $k-1$. For $1 \le k < g$, Form k also has a set of items that are in common with Form $k+1$. For examinees who were administered Form k ($1 \le k < g$), let the random variable U_k refer to the score on the items that are in common with Form $k+1$. (Note that V_1 and U_g are not defined, because Form 1 and Form g are at the ends of the chain.) This data collection design is illustrated in Figure 1. In Figure 1, each column represents an occasion on which a form is administered. The forms are linked together by the common item sets. In Figure 1, lines connect scores corresponding to the same common item set, that is, U_k and V_{k+1} contain the same items.

Figure 1
A Common-Items Nonequivalent Populations Design (X_k = Score on Form k; U_k = Score on Anchor Items Common to Forms k and $k+1$; V_k = Score on Anchor Items Common to Forms k and $k-1$; V_1 and U_g are Empty Sets)

Form/Sample					
1	2	3		$g-1$	g
X_1	X_2	X_3	...	X_{g-1}	X_g
U_1	U_2	U_3	...	U_{g-1}	$*U_g$
$*V_1$	V_2	V_3	...	V_{g-1}	V_g

Each of the g forms is administered to a different sample of examinees. For the examinees administered Form k, let $\mu(X_k)$ be the mean of X_k, $\sigma^2(X_k)$ be the variance of X_k, and $\sigma(X_k, U_k)$ be the covariance of X_k and U_k, and so on. Then estimates of the following eight moments from each sample are needed to perform a linear equating: $\mu(X_k)$, $\sigma^2(X_k)$, $\mu(U_k)$, $\sigma^2(U_k)$, $\mu(V_k)$, $\sigma^2(V_k)$, $\sigma(X_k, U_k)$, and $\sigma(X_k, V_k)$. Because V_1 and U_g do not exist, the moments that involve V_1 and U_g are set to 0.0. Some moments [e.g., $\mu(X_k)$ and $\sigma^2(X_k)$] are shared by two linear functions (e.g., $l_{k-1,k}$ and $l_{k,k+1}$, so estimates of functions $l_{k-1,k}$ and $l_{k,k+1}$ are not statistically independent).

Because the mean and variance of each variable and the covariance between the total score and each common item score are required to compute the equating function, there are eight moments used in each of the g samples for the equatings shown in Figure 1: $\mu(X_k)$, $\sigma^2(X_k)$, $\mu(U_k)$, $\sigma^2(U_k)$, $\mu(V_k)$, $\sigma^2(V_k)$, $\sigma(X_k, U_k)$, and $\sigma(X_k, V_k)$. Let these moments be represented by $\theta_{k1}, \theta_{k2}, \ldots, \theta_{k8}$, and their estimates be represented by $\hat{\theta}_{k1}, \hat{\theta}_{k2}, \ldots, \hat{\theta}_{k8}$, then Equation 3 can be rewritten as:

$$\text{var}\left[\hat{l}_{1,g}\left(x_g\right)\right] \approx \sum_{k=1}^{g}\left[\sum_{i=1}^{8}\left(\frac{\partial l_{1,g}}{\partial \theta_{ki}}\right)^2 \text{var}\left(\hat{\theta}_{ki}\right) + \sum_{i=1}^{8}\sum_{j\neq i=1}^{8} \frac{\partial l_{1,g}}{\partial \theta_{ki}}\frac{\partial l_{1,g}}{\partial \theta_{kj}} \text{cov}\left(\hat{\theta}_{ki}, \hat{\theta}_{kj}\right)\right]. \tag{4}$$

If g (the number of forms) is 2, then Equation 4 is equivalent to Equation 20 in Kolen (1985) and to Equation 31 in Hanson et al. (1993); these latter two equations were used to derive the SEs of Tucker and Levine equating for two forms.

Here, Equation 4 was used to compute the SEs of a chain of equatings using the Levine method (LEM) and Tucker method (TEM) as an example. Formulas for the LEM and TEM are presented by Kolen & Brennan (1987). In their approach, $\mu_s(X_k)$, $\sigma_s^2(X_k)$, $\mu_s(X_{k+1})$, and $\sigma_s^2(X_{k+1})$ are means and variances for a synthetic population that is assumed to be a combination of two populations of potential examinees taking Forms k and $k+1$. Two weights, w_1 and w_2, are used to weight the strata in defining the synthetic population. Examinees administered Form $k+1$ are considered to be a random sample from Population 1, and examinees administered Form k are considered to be a random sample from Population 2. The two populations are proportionally weighted by w_1 and w_2, where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$. Here, w_1 and w_2 are set to be proportional to the sample sizes. More detailed discussion about the synthetic population and weights can be found in Kolen & Brennan (1987).

Computation of partial derivatives. The equating function of a chain of equatings, $l(x)$, is a composite function (note that the subscripts are omitted here for simplicity). Finding the analytic expressions of the partial derivatives of a composite function involves application of the chain rule of differentiation. As the number of forms increases, the analytic expressions of its partial derivatives become more complicated. An easier way to find the partial derivatives of $l(x)$ is to use a numerical approach. Lord (1975) presented an example of using numerical derivatives in computing asymptotic sampling variances. Zeng (1993) described a numerical approach for computing partial derivatives required to compute the SEs of linear equating. Let $\boldsymbol{\theta}$ denote a row vector of variables θ_i, $i = 1, \ldots, s$, where s is the total number of variables used in $l(x)$. Then the first partial derivative of $l(x)$ with respect to θ_i can be approximated by

$$\frac{\partial l(x|\theta)}{\partial \theta_i} = \frac{l(x|\theta+\delta_i) - l(x|\theta-\delta_i)}{2h_i} - O\left(h_i^2\right), \tag{5}$$

where $O\left(h_i^2\right)$ is the error of the approximation, and δ_i is the ith row of the diagonal matrix $\boldsymbol{\Delta}$, where

$$\boldsymbol{\Delta} = \begin{bmatrix} h_1 & & & \\ & h_2 & & \\ & & \ldots & \\ & & & \ldots \quad h_s \end{bmatrix}, \tag{6}$$

and h_i is a small value. Here, h_i was set to $\theta_i/1{,}000$. Because Equation 5 was derived by expanding $l(x|\boldsymbol{\theta})$ at two neighboring points $(\boldsymbol{\theta}+\boldsymbol{\delta}_i)$ and $(\boldsymbol{\theta}-\boldsymbol{\delta}_i)$ with a second order Taylor's series, the magnitude of the error is based on the magnitude of the third partial derivative. If the third partial derivative with respect to θ_i is 0.0, then the first partial derivative with respect to θ_i approximated by Equation 5 is the exact value, otherwise error is involved in the approximation. The error of approximation is bounded by Ch_i^2 where C is the maximum absolute value of the third partial derivative with respect to θ_i. Zeng (1993) showed that the partial derivatives of the TEM computed numerically differ from the exact values by less than 10^{-6}. [More detailed discussion of numerical derivatives with more than one variable can be found in advanced calculus books (e.g., Taylor & Mann, 1983)]. When Equation 5 is used to compute the partial derivatives needed in Equation 4, the equating function l has to be the total $l_{1,g}$ of the chain—the function that converts a score on the scale of Form g, the newest form, to the scale of Form 1, the oldest form.

Computation of the variances and covariances. The computation of the variances and covariances of the moments needed in Equation 4 depends on the nature of the equating chain and the data collection design. In a chain of independent equatings using the random groups design, the moments used in equating functions are obtained from univariate distributions. The computation of variances and covariances of univariate moments is straightforward.

However, in a dependent chain of equatings using the CINPD, one sample of examinees may be used in two equatings. When a sample is used in two equatings, the moments used in the chain equating function are from a trivariate distribution of X, U, and V. In this case, some trivariate moments are involved in the computation of the variances and covariances of the bivariate moments. Two samples, however, are used in the first and last equatings in the chain. For these two samples, only bivariate moments are involved. The variances and covariances of the moments needed in Equation 4 are presented in Table 1. The computations of the variances and covariances that involve univariate and bivariate moments are adapted from Kendall & Stuart (1977, p. 250). The computations of the variances and covariances that involve trivariate moments are derived from these formulas. Because all moments in Table 1 are from one sample, the subscripts are omitted.

Alternative formulas. In summary, the sampling variance of a composite function $\hat{l}_{1,g}(x_g)$ can be computed as follows: (1) find the variances and the appropriate covariances of the moments that are used in $\hat{l}_{1,g}(x_g)$, (2) find the first partial derivatives with respect to all moments evaluated at x_g, and (3) compute $\mathrm{var}\left[\hat{l}_{1,g}(x_g)\right]$ by substituting the first partial derivatives and the variances and covariances into Equation 4.

The method described here is similar to the method described in Lord (1975). The only difference is in the way in which the variances and covariances between the moments are computed. In Lord's method, the variances and covariances are computed assuming a multivariate normal score distribution. The advantage of assuming the multivariate normal distribution is that the computation can be greatly simplified. The disadvantage is that if the score distributions differ considerably from a multivariate normal distribution, the obtained SEs of chain equatings may be biased.

The general delta method (GDM) described in Equation 3 can be used to compute SEs of a chain of linear equatings that used a variety of equating methods. For example, the first link might have used the TEM, the next the LEM, and the last the TEM again. The data collection designs also may vary. For example, the equatings could use a random groups design, a nonequivalent groups internal anchor test design, or a nonequivalent groups external anchor test design.

Consider an alternative version of Equation 3 that is a generalization of one that was suggested by an anonymous reviewer:

$$\mathrm{var}\left[\hat{l}_{1,g}(x_g)\right] \approx \sum_{k=1}^{g}\sum_{m=1}^{s}\left[\sum_{i=1}^{t}\left(\frac{\partial l_{1,g}}{\partial \theta_{kmi}}\right)^2 \mathrm{var}\left(\hat{\theta}_{kmi}\right) + \sum_{i=1}^{t}\sum_{j\neq i=1}^{t}\frac{\partial l_{1,g}}{\partial \theta_{kmi}}\frac{\partial l_{1,g}}{\partial \theta_{kmj}}\mathrm{cov}\left(\hat{\theta}_{kmi},\hat{\theta}_{kmj}\right)\right], \qquad (7)$$

where

Table 1
Sampling Variances and Covariances of Moments of a Trivariate Distribution

Statistic(s)	Sampling Variances and Covariances
$\text{var}[\hat{\mu}(X)]$	$\hat{\sigma}^2(X)/N$
$\text{var}[\hat{\sigma}^2(X)]$	$\{\text{E}[X-\mu(X)]^4-\sigma^4(X)\}/N$
$\text{var}[\hat{\sigma}(X,U)]$	$\{\text{E}[X-\mu(X)]^2[U-\mu(U)]^2-\sigma^2(X,U)\}/N$
$\text{cov}[\hat{\mu}(X),\hat{\mu}(U)]$	$\sigma(X,U)/N$
$\text{cov}[\hat{\mu}(X),\hat{\sigma}^2(X)]$	$\text{E}[X-\mu(X)]^3/N$
$\text{cov}[\hat{\mu}(X),\hat{\sigma}^2(U)]$	$\text{E}[X-\mu(X)][U-\mu(U)]^2/N$
$\text{cov}[\hat{\mu}(X),\hat{\sigma}(X,U)]$	$\text{E}[X-\mu(X)]^2[U-\mu(U)]/N$
$\text{cov}[\hat{\sigma}^2(X),\hat{\sigma}^2(U)]$	$\{\text{E}[X-\mu(X)]^2[U-\mu(U)]^2-\sigma^2(X)\sigma^2(U)\}/N$
$\text{cov}[\hat{\sigma}^2(U),\hat{\sigma}(X,U)]$	$\{\text{E}[X-\mu(X)]^3[U-\mu(U)]-\sigma^2(X)\sigma(X,U)\}/N$
$\text{cov}[\hat{\mu}(V),\hat{\sigma}(X,U)]$	$\{\text{E}[V-\mu(V)][X-\mu(X)][U-\mu(U)]\}/N$
$\text{cov}[\hat{\sigma}^2(V),\hat{\sigma}(X,U)]$	$\{\text{E}[V-\mu(V)]^2[X-\mu(X)][U-\mu(U)]-\sigma^2(V)\sigma(X,U)\}/N$
$\text{cov}[\hat{\sigma}(X,V),\hat{\sigma}(X,U)]$	$\{\text{E}[X-\mu(X)]^2[V-\mu(V)][U-\mu(U)]-\sigma(X,V)\sigma(X,U)\}/N$

$\hat{\theta}_{kmi}$ is the estimated moment i of sample m of Form k,
g is the number of forms,
s is the number of samples used in link k ($s=2$ for $1<k<g$, else $s=1$), and
t is the number of moments used in computing the equating function in each link.

Equation 7 can be used to display chains that mix equating methods and data collection designs.

Suppose, for example, that interest is in a CINPD equating chain consisting of g forms, each of which (except for the first and the last) contains two different external equating sections. In this case, Equation 7 can be applied by setting $t=5$ to represent 5 moments for each sample—means and variances of the equating and nonequating items and the covariance between the equating and nonequating items. The more general form of Equation 7 also could be used to represent a chain of g linear equatings using the random groups design. In this case $t=2$, which represents the mean and variance of each form score for each sample.

Empirical Study

Method

Data. The behavior of the SEs of a chain of linear equatings was studied using data from three forms of a 125-item test. Form 1 was administered to 795 examinees from Population 1, Form 2 was administered to 1,793 examinees from Population 2, and Form 3 was administered to 773 examinees from Population 3. Form 2 was administered one year after Form 1, and Form 3 was administered one year after Form 2.

The random variable U_1 represented the score on the 30 common items in Form 1, the random variable G_1 represented the remaining 95 items in Form 1, and $X_1=U_1+G_1$. Form 2 contained a set of 30 items that were in common with Form 1 and a set of 30 items that were in common with Form 3. For Form 2, the random variable V_2 represented the score on the 30 items in common with Form 1, the random variable U_2 represented the score on the 30 items in common with Form 3, the random variable G_2 represented the remaining 65 items in Form 2, and $X_2=G_2+U_2+V_2$. Form 3 contained a set of 30 items in common with Form 2. For Form 3, the random variable V_3 represented the score on the 30 common items in Form 3, the random variable G_3 represented the remaining 95 items, and $X_3=G_3+V_3$. The chain of equatings was X_3

equated to X_2, which in turn was equated to X_1.

Distributions. A four-parameter beta-binomial model was used to fit the bivariate distributions (G_1, U_1) and (G_3, V_3), and the trivariate distribution (G_2, U_2, V_2) using the procedure described in Lord (1965). The fitted distributions were used as population distributions for generating random samples. The steps used to fit the trivariate distribution of G_2, U_2, and V_2 are presented below. Analogous procedures were used to fit the bivariate distributions (G_1, U_1) and (G_3, V_3).

1. A four-parameter beta binomial model was fit to the univariate distributions of G_2, U_2, and V_2 using the procedure described in Hanson (1991). This resulted in estimates of the true score distributions corresponding to G_2, U_2, and V_2 in Population 2. Let $F_G(\tau)$, $F_U(\tau)$, and $F_V(\tau)$ denote the cumulative distribution functions of the proportion-correct true score corresponding to G_2, U_2, and V_2, respectively, in Population 2.
2. Assuming that the true scores corresponding to G_2, U_2, and V_2 were related by monotonically increasing functions and G_2, U_2, and V_2 were conditionally independent given true score, the joint distribution of G_2, U_2, and V_2 in Population 2 is given by:

$$P(G_2=i,V_2=j,U_2=k)=\int_l^u P(G_2=i|\tau)P[V_2=j|\psi(\tau)]P[U_2=k|\eta(\tau)]g(\tau)d(\tau), \tag{8}$$

where

$$\psi(\tau)=F_V^{-1}[F_G(\tau)], \eta(\tau)=F_U^{-1}[F_G(\tau)], \tag{9}$$

$g(\tau)$ is the density function of the proportion-correct true score corresponding to G_2 in Population 2, and $g(\tau)>0$ for $l \le \tau \le u$, where l and u are the lower and upper limits, respectively, of the proportion-correct true score distribution. Under the beta binomial model, the error distributions [e.g., $P(G_2=i|\tau)$] are binomial. To estimate the joint distribution of G_2, U_2, and V_2, estimates of the true score distributions produced in Step 1 were substituted in Equation 8. The integral in Equation 8 was evaluated using 64-point Gauss-Legendre quadrature (Thisted, 1988).

Estimating SEEq by simulation. Estimates of the SEEq X_3 to X_1 were computed for several sample sizes (N) by simulation using the fitted distributions of (G_1, U_1), (G_2, V_2, U_2), and (G_3, V_3) as population distributions. The simulations were carried out as follows:

1. Samples of size N were drawn from the bivariate distributions of (G_1, U_1) and (G_3, V_3) and the trivariate distribution of (G_2, V_2, U_2).
2. Using the data from Step 1, the linear function equating X_3 to X_1 through X_2 was calculated using both the TEM and LEM. The SEEqs were calculated using Equation 4, which matched the actual data collection design used, and Equation 2, which assumed that the equatings in the chain were statistically independent.
3. Steps 1 and 2 were repeated 10,000 times. It was found that 10,000 replications were sufficient to produce very stable estimates.

The SEEq for a particular equating method at $X_3 = x_3$ was estimated as:

$$\text{SEEq}\left[\hat{l}_{1,3}(x_3)\right]=\left\{\frac{\sum_{i=1}^{10,000}\left[\hat{l}_{1,3,i}(x_3)-\bar{l}_{1,3}(x_3)\right]^2}{10,000}\right\}, \tag{10}$$

where

$$\bar{l}_{1,3}(x_3)=\frac{1}{10,000}\sum_{i=1}^{10,000}\hat{l}_{1,3,i}(x_3), \tag{11}$$

and $\hat{l}_{1,3,i}(x_3)$ is the estimated equating function for replication i. The SEEq given by Equation 10 was treated as the true SEEq and the bias of a particular estimated SEEq at $X_3 = x_3$ was estimated by

$$\frac{1}{10{,}000}\sum_{i=1}^{10{,}000}\left\{\widehat{\text{SEEq}}_i\left[\hat{l}_{1,3}(x_3)\right]\right\} - \text{SEEq}\left[\hat{l}_{1,3}(x_3)\right], \qquad (12)$$

where $\widehat{\text{SEEq}}_i\left[\hat{l}_{1,3}(x_3)\right]$ is the estimated SE for replication i. The simulations were conducted using N = 250, N = 500, and N = 1,000 examinees.

Results and Discussion

The first four central moments of the fitted score distributions for the three achievement test forms are reported in Table 2. The first four central moments of the fitted score distributions were the same as the corresponding moments for the observed score distributions. The three test forms were relatively easy—over 75% of the items were answered correctly. [This was calculated by summing the mean of noncommon items (G) and means of common items (U and R) for each form, and dividing by the total number of items (125).] The score distributions were considerably negatively skewed and had kurtosis indexes higher than that for a normal distribution. The correlation coefficients between the noncommon and common item scores for the fitted and observed score distributions are presented in Table 3. Based on Tables 2 and 3, the fitted score distributions appear to be reasonably similar to the observed score distributions. These fitted score distributions were used as population distributions in the simulation.

Table 2
Mean, Standard Deviation (SD), Skewness, and Kurtosis of the Fitted Distributions for Three Forms of an Achievement Test

Variable	Mean	SD	Skewness	Kurtosis
G_1	73.886	9.830	−.944	3.867
U_1	22.955	4.155	−.909	3.631
G_2	51.209	5.820	−1.113	5.297
U_2	23.009	4.265	−.859	3.714
V_2	22.828	3.789	−.776	3.451
G_3	72.339	10.017	−.983	3.906
V_3	23.406	3.964	−.931	3.528

The bias of the SE of $\hat{l}_{1,3}(x_3)$ computed with the GDM (Equation 4) and the method assuming independent equatings (IEM; Equation 2) for each selected score point are reported in Table 4. The results of the simulation indicated that for a particular number-correct score, the SE of a chain of linear equatings decreased as N increased. For example, for a number-correct score of 70, the SEEq for the LEM decreased from 3.564 (N = 250) to 1.760 (N = 1,000). This trend is consistent with the notion that if N is infinitely large, then the obtained equivalent is the same as the population value and the expected SEEq is 0.0. For a fixed N, the SEs were smaller for scores near the mean than for scores at the two ends of the score range. For example, in the case of N = 250, the SEEq for the LEM was 1.012 for a number-correct score of 100, which is close to the mean score of 96. This SEEq was smaller than those for the other number-correct scores further from the mean. The SEs for the LEM were generally larger than those for the TEM.

The bias of the SE of $\hat{l}_{1,3}(x_3)$ computed with the GDM and the IEM for each selected score point also is reported in Table 4. These bias indexes were computed by subtracting the true values from the average of the estimated SEs obtained over the 10,000 replications.

For the LEM, the estimated SEs given by the IEM were considerably less accurate than those given by the

Table 3
Correlations Between Noncommon and Common Item Scores For the Fitted and Observed Score Distributions

Variables	Fitted	Observed
G_1U_1	.78	.80
G_2U_2	.70	.63
G_2V_2	.66	.63
G_3V_3	.76	.80

GDM, which does not require such an assumption. For example, for $N = 1{,}000$ the absolute value of the bias of the estimated SEs of the LEM ranged from .001 to .014 for the GDM and from .032 to .107 for the IEM. The IEM underestimated the SEs of the LEM for all N. The same patterns can be observed for $N = 250$ and $N = 500$. The bias of the estimates for both methods decreased as N increased.

For the TEM, the absolute value of the bias of estimated SEs from the GDM ranged from .005 to .055 for $N = 250$, from .002 to .040 for $N = 500$, and from .001 to .005 for $N = 1{,}000$. The absolute value of the bias of estimated SEs given by the IEM ranged from .004 to .252 for $N = 250$, from .004 to .180 for $N = 500$, and from .002 to .162 for $N = 1{,}000$. The IEM tended to overestimate the SEs of the TEM. For example, for $N = 250$ and a number-correct score of 70, the estimated SE by the IEM was .252 larger than the true value.

The example used in the simulations was a chain of two linear equatings that linked three forms. The GDM can be applied easily to chains that link more than three forms. Because there are no iterations involved in computing the numerical derivatives needed in Equation 2, the entire process for computing SEs

Table 4
Results of Simulations Based on 10,000 Replications for the GDM and IEM for Scores of 70 to 120 and $N = 250$, 500, and 1,000 Using the LEM and TEM

			LEM			TEM		
	$\hat{l}_{1,3}(x_3)$			$\widehat{\text{SEEq}}-\text{SEEq}$			$\widehat{\text{SEEq}}-\text{SEEq}$	
N and x_3	LEM	TEM	SEEq	GDM	IEM	SEEq	GDM	IEM
$N = 250$								
70	61.25	64.55	3.564	−.084	−.271	2.332	−.055	.252
80	73.65	75.87	2.532	−.054	−.192	1.696	−.033	.163
90	86.05	87.19	1.587	−.023	−.117	1.125	−.010	.072
100	98.44	98.50	1.012	.003	−.064	.777	.005	−.004
110	110.84	109.82	1.390	−.013	−.088	.944	−.010	.035
120	123.24	121.14	2.289	−.043	−.156	1.459	−.035	.127
$N = 500$								
70	61.25	64.55	2.514	−.046	−.179	1.658	−.040	.180
80	73.65	75.87	1.786	−.030	−.128	1.207	−.025	.115
90	86.05	87.19	1.121	−.013	−.080	.800	−.010	.048
100	98.44	98.50	.718	−.001	−.048	.551	.002	−.004
110	110.84	109.82	.984	−.010	−.063	.667	−.005	.028
120	123.24	121.14	1.617	−.026	−.106	1.032	−.021	.096
$N = 1{,}000$								
70	61.25	64.55	1.760	−.014	−.107	1.142	.005	.162
80	73.65	75.87	1.252	−.010	−.079	.833	.004	.105
90	86.05	87.19	.788	−.005	−.052	.556	.004	.045
100	98.44	98.50	.506	.001	−.032	.388	.003	−.002
110	110.84	109.82	.688	.001	−.036	.467	.002	.025
120	123.24	121.14	1.128	.001	−.059	.715	.001	.085

of a chain of linear equatings is fast. Based on the implementation on a Macintosh IIcx microcomputer (with a math coprocessor), only a few seconds of time were needed to compute the SEs of a chain for equating three forms with $N = 1{,}000$.

Conclusions

The GDM was reasonably accurate in estimating the SEs of a chain of linear equatings under the CINPD for the example studied. The SEs given by the IEM were less accurate than those given by the GDM.

References

American College Testing Program. (1988). *ACT assessment program technical manual.* Iowa City IA: Author.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton NJ: Educational Testing Service.

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. 9–40). New York: Academic Press.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Report 91-5). Iowa City IA: American College Testing Program.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17,* 225–237.

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

Kolen, M. J. (1985) Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209–223.

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

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

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

Lord, F. M. (1975). Automated hypothesis tests and standard errors for nonstandard problems. *The American Statistician, 29,* 56–59.

Taylor, A. E., & Mann, W. R. (1983). *Advanced calculus* (3rd ed.). New York: Wiley.

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

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 17,* 177–186.

Acknowledgments

The authors thank two anonymous reviewers and Ronald Cope for their comments and suggestions in revising the manuscript.

Author's Address

Send requests for reprints or further information to Lingjia Zeng, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.