An Alternative Approach for IRT Observed-Score Equating of Number-Correct Scores

Lingjia Zeng and Michael J. Kolen
American College Testing

An alternative approach for item response theory observed-score equating is described. The number-correct score distributions needed in equating are found by numerical integration over the theoretical or empirical distributions of examinees' traits. The item response theory true-score equating method and the observed-score equating method described by Lord, in which the number-correct score distributions are summed over a sample of trait estimates, are compared in a real test example. In a computer simulation, the observed-score equating methods based on numerical integration and summation were compared using data generated from standard normal and skewed populations. The method based on numerical integration was found to be less biased, especially at the two ends of the score distribution. This method can be implemented without the need to estimate trait level for individual examinees, and it is less computationally intensive than the method based on summation. *Index terms: equating, item response theory, numerical integration, observed-score equating.*

In item response theory (IRT) observed-score (OS) equating of number-correct (NC) scores, an appropriate IRT model is used to produce an estimated distribution of observed NC scores (ONCSs) on each test form to be equated. Conventional linear or equipercentile equating is then conducted on these estimated OS distributions. The goal of IRT OS equating is for the distributions of equated NC scores for the test forms to be as similar as possible.

Lord (1982) described an implementation of IRT OS equating in which the OS distribution is first generated for an individual of a given estimated trait level (θ), and then cumulated over a sample of individuals. Because the estimated θ of each individual is used to generate the OS distribution, Lord suggested that the errors resulting from θ estimation might result in systematic error in the equating process. Another drawback of the procedure described by Lord is that it is very time consuming. However, with a reasonable estimate of the distribution of θ, the population OS distribution can be easily approximated numerically. If a marginal maximum likelihood approach (e.g., BILOG; Mislevy & Bock, 1990) is used to estimate item parameters, a posterior distribution of θ in quadrature form is available and can be used to generate the OS distribution.

The primary purpose of this paper was to describe and evaluate an alternative approach for IRT OS equating in which the NC score distributions are integrated over theoretical or empirical distributions of examinee θs. Comparisons with alternate implementations of Lord's method were made using a real test example and a computer simulation. In addition, in the real test example IRT OS equating was compared to IRT true-score equating (Lord, 1982), in which the true NC scores are equated through the test response functions for the two forms.

Method

Consider two forms, Form X and Form Y, of a test of K items, administered to two randomly equivalent samples drawn from a population of examinees. Let $\boldsymbol{\omega}_i$ denote a vector of item parameters under an appropriate IRT model for item i on Form X. Let θ denote a random variable of trait level drawn from a population of examinees. Then the probability of answering the ith item on Form X correctly can be expressed as $p(\theta, \boldsymbol{\omega}_i)$. Similarly, let $\boldsymbol{\lambda}_i$ denote a vector

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 3, September 1995, pp. 231–240

0146-6216/95/030231-10$1.75

of item parameters for the ith item on Form Y. The probability of answering item i on Form Y correctly is $p(\theta, \boldsymbol{\lambda}_i)$. Let X ($X = 0, 1, \ldots, K$) denote a random variable of a NC score on Form X. Assuming local independence, $f(X|\theta)$, the ONCS distribution given θ can be generated by the recursion formula described by Lord & Wingersky (1984). Suppose θ has a known distribution with the density function $g(\theta)$. Then $f(X)$, the ONCS distribution on Form X for the population, can be found by

$$f(X) = \int_{-\infty}^{\infty} f(X|\theta) g(\theta) d(\theta). \quad (1)$$

This integral can be approximated numerically to any specified degree of precision. Similarly, the population ONCS distribution on Form Y, $f(Y = 0, 1, \ldots, K)$ where Y is a random variable for a score on Form Y, can be found by

$$f(Y) = \int_{-\infty}^{\infty} f(Y|\theta) g(\theta) d(\theta). \quad (2)$$

With $f(X)$ and $f(Y)$, conventional linear or equipercentile equating methods can be applied to produce the equating relationship.

The ONCS distributions $f(X)$ and $f(Y)$ must be computed from the population item parameters, which are not available in practice. The estimates of the ONCS distributions can be found using the estimated item parameters. Let $\hat{\boldsymbol{\omega}}$ be a $K \times m$ matrix (where m is the number of parameters in the IRT model) of estimated item parameters for Form X obtained from the random sample of people taking Form X, and $\hat{\boldsymbol{\lambda}}$ be a $K \times m$ matrix of the estimated item parameters for Form Y obtained from the random sample of people taking Form Y. A smoothed version of the ONCS distribution of Form X given $\hat{\boldsymbol{\omega}}$ in Sample 1, the examinees taking Form X, can be found by

$$\hat{f}_1(X|\hat{\boldsymbol{\omega}}) = \int_{-\infty}^{\infty} \hat{f}(X|\hat{\boldsymbol{\omega}}, \theta) \hat{g}_1(\theta) d(\theta), \quad (3)$$

where $\hat{g}_1(\theta)$ is the estimated density function of θ obtained using Sample 1 for the population of potential examinees who would take the test. The smoothed version of the ONCS distribution of Form Y given $\hat{\boldsymbol{\lambda}}$ in Sample 2, the examinees taking Form Y, can be found by

$$\hat{f}_2(Y|\hat{\boldsymbol{\lambda}}) = \int_{-\infty}^{\infty} \hat{f}(Y|\hat{\boldsymbol{\lambda}}, \theta) \hat{g}_2(\theta) d(\theta), \quad (4)$$

where $\hat{g}_2(\theta)$ is the estimated density function of θ obtained using Sample 2 for the population of potential examinees who would take the test. The estimated score distribution of Form X given $\hat{\boldsymbol{\omega}}$ using Sample 2 can be found by

$$\hat{f}_2(X|\hat{\boldsymbol{\omega}}) = \int_{-\infty}^{\infty} \hat{f}(X|\hat{\boldsymbol{\omega}}, \theta) \hat{g}_2(\theta) d(\theta). \quad (5)$$

The estimated score distribution of Form Y given $\hat{\boldsymbol{\lambda}}$ using Sample 1 can be found by

$$\hat{f}_1(Y|\hat{\boldsymbol{\lambda}}) = \int_{-\infty}^{\infty} \hat{f}(Y|\hat{\boldsymbol{\lambda}}, \theta) \hat{g}_1(\theta) d(\theta). \quad (6)$$

The estimated score distribution of Form X given $\hat{\boldsymbol{\omega}}$ using Sample S, a synthetic sample of people taking both forms, can be found by

$$\hat{f}_S(X|\hat{\boldsymbol{\omega}}) = w_1 \hat{f}_1(X|\hat{\boldsymbol{\omega}}) + w_2 \hat{f}_2(X|\hat{\boldsymbol{\omega}}), \quad (7)$$

where w_1 and w_2 are weights used to weight the strata in defining the synthetic population. The synthetic population is conceived of as containing two strata, Stratum 1 and Stratum 2. Examinees administered Form X are considered to be a random sample from Stratum 1; examinees administered Form Y are considered to be a random sample from Stratum 2. In a random groups design, the two strata are considered equivalent. That is, Sample 1 and Sample 2 are equivalent random samples from the same population. In the common-item nonequivalent-groups design, the two strata are considered to represent two nonequivalent populations. The two strata can be proportionally weighted by w_1 and w_2, where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$. More detailed discussion about the synthetic population and weights can be found in Kolen & Brennan (1987, 1995). The estimated score distribution of Form Y given $\hat{\boldsymbol{\lambda}}$ in Sample S can be found by

$$\hat{f}_S(Y|\hat{\boldsymbol{\lambda}}) = w_1 \hat{f}_1(Y|\hat{\boldsymbol{\lambda}}) + w_2 \hat{f}_2(Y|\hat{\boldsymbol{\lambda}}). \quad (8)$$

With $\hat{f}_S(X|\hat{\boldsymbol{\omega}})$ (Equation 7) and $\hat{f}_S(Y|\hat{\boldsymbol{\lambda}})$ (Equation 8), a conventional equipercentile or linear equating

method can be applied to estimate the equating relationship.

The method described here can be applied in either the randomly equivalent groups or the common-item nonequivalent-groups equating design. For the common-item nonequivalent-groups equating design, $\hat{g}_1(\theta)$ and $\hat{g}_2(\theta)$ are estimates of the density functions for the two different populations of examinees. In the case of the randomly equivalent groups equating design, the density functions $\hat{g}_1(\theta)$ and $\hat{g}_2(\theta)$ are considered as two estimates for the same population. In practice, when a marginal maximum likelihood approach (e.g., BILOG; Mislevy & Bock, 1990) is used to estimate item parameters, a posterior distribution of θ in quadrature form is available. This posterior distribution can be used to estimate the ONCS distribution by approximating the integrals in Equations 3–6.

Lord (1982) described a method to estimate an ONCS distribution by summing the NC score distribution given an estimate of θ over a sample of examinees. Lord's method can be expressed as

$$\hat{f}(X|\hat{\boldsymbol{\omega}}) = \sum_{j=1}^{N} \hat{f}_j\left(X|\hat{\boldsymbol{\omega}}, \hat{\theta}_j\right), \tag{9}$$

where $\hat{f}_j\left(X|\hat{\boldsymbol{\omega}}, \hat{\theta}_j\right)$ is the estimated NC score distribution given $\hat{\theta}_j$, the estimated θ level of examinee j, and N is the sample size. Lord's method is referred to here as the summation method.

Real-Data Example

Method

The data used in this example were obtained from administering two forms of an American College Testing (ACT) mathematics test (American College Testing Program, 1989), Form X and Form Y, to two randomly equivalent groups of examinees. Form X was administered to 2,800 examinees; Form Y was administered to 2,903 examinees. The test contained 60 multiple-choice items. The NC score distributions of the two forms are shown in Figure 1. The distributions of both forms were slightly positively skewed. The Form X distribution appears to be more peaked than that of Form Y.

The item parameters were estimated using BILOG (Mislevy & Bock, 1990). The prior θ distributions were assumed to be normal and were approximated with 40 equally spaced quadrature points. The estimated empirical θ probability distributions (posterior distributions) computed by BILOG for the two forms are plotted in Figure 2. The estimated distributions appear to be symmetric and bell-shaped. The estimated θ distribution for examinees taking Form X appears to be slightly more peaked than that for examinees taking Form Y.

Figure 1
Number-Correct Score Distributions for Form X and Form Y

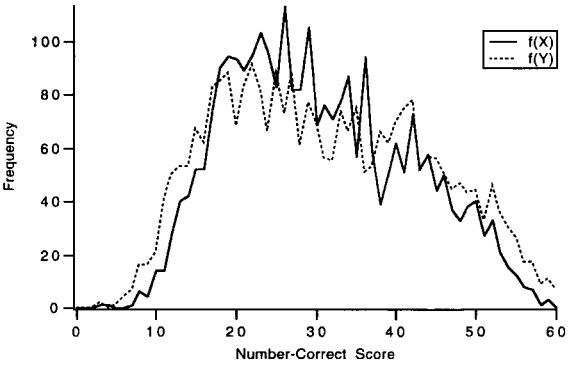
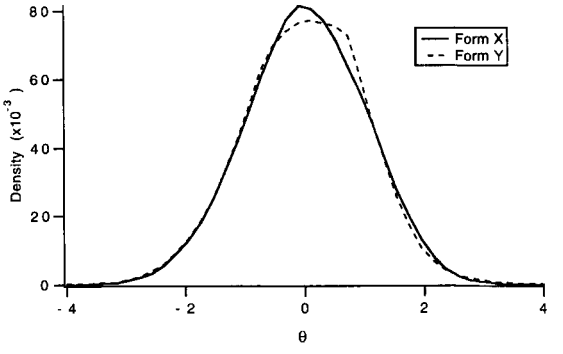


Figure 2
Estimated θ Distributions for Form X and Form Y



Estimates of item parameters were used to estimate the ONCS distributions using the numerical integration method (Equations 7 and 8) and the summation method (Equation 9). For the integration method, posterior distributions of θ provided by

BILOG were used to replace the density functions in Equations 3–6. The synthetic population was formed using equal weights ($w_1 = w_2 = .5$). For the summation method, maximum likelihood (ML) estimates of θ were used. For comparison purposes, results for IRT true-score equating and unsmoothed equipercentile equating are also provided.

Results

The moments of the NC score, estimated NC score, and the equated score distributions are presented in Table 1. In general, the moments of the estimated score distributions of Form X and Form Y computed using the numerical integration method were reasonably close to those of the NC score distributions. The standard deviations (SDs) of the estimated score distributions and equated score distributions computed using the summation method were considerably larger than those for the NC score distributions, the estimated score distributions, and the equated score distribution computed using the integration method. For example, the SD of the equated score distribution was 12.7575 for the integration method and 13.1452 for the summation method. This result agrees with Han's (1993) study in which she found that the SDs of the score distributions estimated with ML estimates were larger than those of the OS distributions. The large SD is caused by large errors in the ML estimates (Mislevy, 1993).

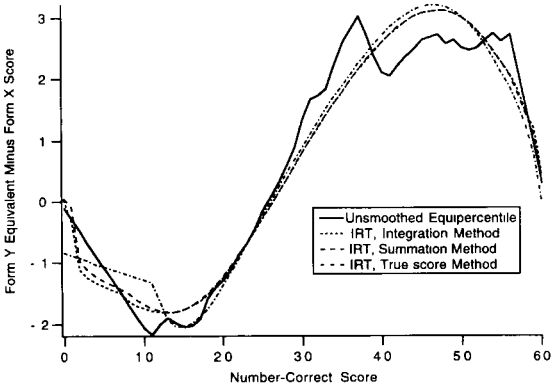
IRT true-score equating (Lord, 1982) is an often-used alternative for equating NC scores, which is why it was included here. In this procedure, true scores are equated using the test response functions for the two forms. Lord & Wingersky (1984) found very similar results for IRT true and OS equating, whereas Kolen (1981), Han (1993), and Kolen & Brennan (1995) indicated that the methods tend to produce somewhat different results. The four sets of Form Y equivalents computed using the unsmoothed equipercentile equating method, the integration and sum-

Table 1
Moments of the NC Score, Estimated NC Score, and Equated Score Distributions for an ACT Mathematics Test

Distribution	Mean	SD	Skewness	Kurtosis
NC Score (X)	30.3886	11.1084	.3236	2.2527
NC Score (Y)	31.0823	12.7221	.2078	2.0597
Score Distributions Estimated Using Integration Method				
$\hat{f}_1(X\|\hat{\boldsymbol{\omega}})$	30.3901	11.1195	.3329	2.2607
$\hat{f}_2(X\|\hat{\boldsymbol{\omega}})$	30.4418	11.0646	.3007	2.2479
$\hat{f}_S(X\|\hat{\boldsymbol{\omega}})$	30.4159	11.0921	.3169	2.2542
$\hat{f}_2(Y\|\hat{\boldsymbol{\lambda}})$	31.0893	12.7105	.2305	2.0543
$\hat{f}_1(Y\|\hat{\boldsymbol{\lambda}})$	31.0176	12.7599	.2646	2.0751
$\hat{f}_S(Y\|\hat{\boldsymbol{\lambda}})$	31.0534	12.7353	.2476	2.0646
Score Distributions Estimated Using Summation Method				
$\hat{f}_1(X\|\hat{\boldsymbol{\omega}})$	30.4425	11.5049	.2861	2.2439
$\hat{f}_2(X\|\hat{\boldsymbol{\omega}})$	30.4641	11.4264	.2608	2.2643
$\hat{f}_S(X\|\hat{\boldsymbol{\omega}})$	30.4533	11.4657	.2735	2.2540
$\hat{f}_2(Y\|\hat{\boldsymbol{\lambda}})$	31.1240	13.0633	.1919	2.0552
$\hat{f}_1(Y\|\hat{\boldsymbol{\lambda}})$	31.0845	13.1503	.2213	2.0576
$\hat{f}_S(Y\|\hat{\boldsymbol{\lambda}})$	31.1043	13.1069	.2067	2.0564
Equated Score Distributions				
$\hat{e}_Y(X)$, Unsmoothed	31.0799	12.7138	.2063	2.0562
$\hat{e}_{YS}(X)$, Integration	31.0185	12.7575	.2644	2.0734
$\hat{e}_{YS}(X)$, Summation	31.0869	13.1452	.2217	2.0556
$\hat{e}_{YS}(X)$, True Score	31.0379	12.8060	.2476	2.0433

mation IRT OS equating methods, and IRT true-score equating method are plotted in Figure 3. Consistent with Kolen & Brennan and Han, Figure 3 shows that the true-score method produced results similar to those for OS equating, except at very high scores and near the sum of the pseudo-guessing (c) parameter estimates (below which true scores do not exist) for the two forms. Figure 3 shows that the two equating functions obtained by the integration and summation IRT OS equating methods were very similar throughout the NC score range.

Figure 3
Four Equating Functions for the ACT Mathematics Test



Computer Simulations

Method

Two computer simulations were conducted to evaluate the accuracy of the proposed IRT OS equating method. In the first simulation, the θs for the simulated examinees were generated from a standard normal (SN) distribution with mean 0 and unit SD. In the second simulation, the θs for the simulees were generated from a negatively skewed distribution with mean = .1, SD = 1.1, skewness = −.3, and kurtosis = 2.5.

To make the simulations realistic, the real-data ACT mathematics test example was modeled in the simulations. The item parameter estimates were used as the population item parameters in the simulations. The simulation was conducted using three samples sizes: N = 500, 1,000, and 2,000. For a simulated examinee with trait level θ, the 0-1 score on item i (i = 1 to 60) was generated in the following manner:

1. Compute $p_i(\theta)$, the probability of a correct response to the ith item, using the three-parameter logistic model (Birnbaum, 1968);
2. Generate a random number, r, from a uniform distribution between 0 and 1;
3. Assign a 1 (a correct score) if $p_i(\theta) > r$, otherwise assign a 0 (an incorrect score).

The simulation was conducted using the following steps:

Step 1. An N-examinee × 60-item matrix of response scores was generated for Form X and for Form Y.

Step 2. BILOG was run to estimate item parameters, θ distributions, and θ estimates ($\hat{\theta}$s). The $\hat{\theta}$s were estimated using ML, expected a posteriori (EAP), and Bayes modal (BM) estimation.

Step 3. Observed NC score equating was performed to equate Form X to Form Y. First, the estimated NC score distributions for Forms X and Y were computed using the equally weighted synthetic populations (Equations 7 and 8). Then equipercentile equating was performed using the estimated NC score distributions. The score distributions needed in Equations 7 and 8 were computed using the integration method described here and the summation method described by Lord (1982). For the integration method, the population distributions and the posterior θ distributions obtained from BILOG were used. For the summation method, the three types of $\hat{\theta}$s were used.

Steps 1–3 were replicated 50 times. The estimated Form Y equivalent of the Form X score, $\hat{e}_Y(x)$, was compared with the population equivalent, $e_Y(x)$. $e_Y(x)$ was computed using the item parameters for the two forms and the population θ distribution. The mean square error (MSE) for score point x for a particular method was defined by

$$\mathrm{MSE}\left[\hat{e}_Y(x)\right] = \frac{1}{50}\sum_{k=1}^{50}\left[\hat{e}_{Yk}(x) - e_Y(x)\right]^2, \qquad (10)$$

where $\hat{e}_{Yk}(x)$ is $\hat{e}_Y(x)$ for the kth replication. MSE(x) consists of the variance of $\hat{e}_Y(x)$ and the squared bias (SB) of $\hat{e}_Y(x)$. The variance was defined as

$$\mathrm{Var}[\hat{e}_Y(x)] = \frac{1}{50}\sum_{k=1}^{50}[\hat{e}_{Yk}(x) - \bar{e}_Y(x)]^2, \quad (11)$$

where

$$\bar{e}_Y(x) = \frac{1}{50}\sum_{k=1}^{50}\hat{e}_{Yk}(x). \quad (12)$$

SB was defined as

$$\mathrm{SB}[\hat{e}_Y(x)] = [\bar{e}_Y(x) - e_Y(x)]^2. \quad (13)$$

Average MSE (AMSE) over all 60 score points was computed as

$$\mathrm{AMSE} = \sum_{x=1}^{60}\mathrm{MSE}[\hat{e}_Y(x)]P(x), \quad (14)$$

where $P(x)$ is the population proportion of simulated examinees who have an OS of x on Form X. The variance over all 60 score points was computed as

$$\mathrm{Var} = \sum_{x=1}^{60}\mathrm{Var}[\hat{e}_Y(x)]P(x). \quad (15)$$

The SB over all 60 score points was computed as

$$\mathrm{SB} = \mathrm{AMSE} - \mathrm{Var}. \quad (16)$$

Results

The standard normal population. The AMSE (Equation 14), variance (Equation 15), and SB (Equation 16) for the SN population are presented in Table 2. These error indexes represent equating errors for five sets of equating results: two for the integration method and three for the summation method.

In order to use the integration method for IRT OS equating, the density function of the θ distribution must be assumed or estimated. In the simulation with the SN population, the SN distribution was used as a theoretical θ distribution, and the posterior θ distribution given by BILOG was used as an empirical θ probability distribution. As indicated by the AMSEs in Table 2, the equating results obtained by the integration method using the two assumed distributions were very similar to each other and slightly better than those of the summation method.

In IRT OS equating using the summation method, the $\hat{\theta}$s of a pair of samples of examinees are used to generate the OS distributions. Different types of $\hat{\theta}$s may result in different equating results when the summation method is used. In this study, the three types of $\hat{\theta}$s provided by BILOG were investigated. Table 2 shows that the difference among the AMSEs for the three sets of equating results obtained from the summation method were larger than the differences between the AMSEs for the two sets of equating results obtained from the integration method.

The AMSE indexes in Table 2 show that the equating errors for the EAP estimates were smaller than those for the ML and BM estimates. Table 2 also shows that the error variances accounted for a large proportion of the AMSEs. The error variances decreased as the sample size became larger. The SB was generally very small (<.1) for the SN population.

MSE and SB at each integer NC score point for the SN population and $N = 2{,}000$ are plotted in Figures 4a and 4b, respectively. Figure 4a indicates that the MSE plots for the equating equivalents obtained from the integration method appear to be smooth and very similar to each other. The MSE plots for the three summation method equating equivalents were less smooth at the two extremes of the score range. The MSEs for the three summation method curves were larger than those of the integration method at the lower and upper ends of the score distribution.

The SB plots in Figure 4b indicate that the SBs for the three summation methods, and especially for ML $\hat{\theta}$, were considerably larger than those for the two integration methods at the two extremes of the score range. The bumps for the posterior distribution integration method in the 30–40 range of NC scores might be because only 40 quadrature points were used, although the effect on the MSE was minimal.

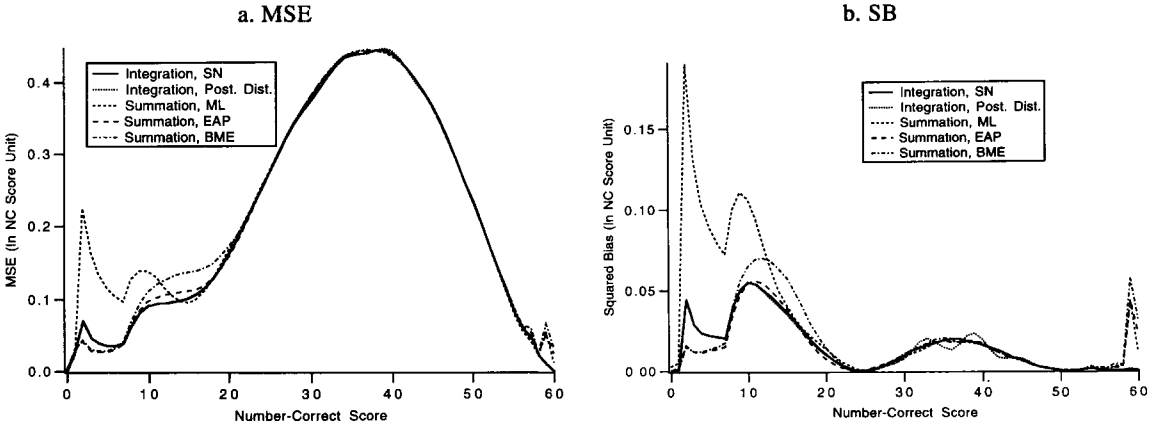
The skewed population. The AMSE, variance, and SB indexes for this population also are presented in Table 2. Equating errors were larger for this population than they were for the SN population. The SB indexes were relatively large for all three sample sizes. The error variances were also large, but they decreased as the sample size increased. Comparing

Table 2
Equating Errors for the Standard Normal and Skewed Population Simulations

Population, *N*, and Error Type	Integration Method		Summation Method		
	Population Distribution	Posterior Distribution	ML	EAP	BM
Standard Normal					
N = 500					
AMSE	.9720	.9721	.9739	.9722	.9834
Variance	.8804	.8805	.8759	.8821	.8840
SB	.0915	.0915	.0980	.0900	.0994
N = 1,000					
AMSE	.4095	.4095	.4115	.4098	.4150
Variance	.3755	.3762	.3753	.3767	.3774
SB	.0340	.0333	.0362	.0331	.0376
N = 2,000					
AMSE	.2863	.2861	.2870	.2866	.2919
Variance	.2741	.2744	.2733	.2748	.2765
SB	.0122	.0117	.0136	.0118	.0154
Skewed					
N = 500					
AMSE	1.3760	1.3870	1.3943	1.3857	1.3948
Variance	1.1252	1.1220	1.1194	1.1265	1.1326
SB	.2508	.2650	.2749	.2592	.2622
N = 1,000					
AMSE	1.1802	1.1906	1.1990	1.1858	1.1968
Variance	.9274	.9238	.9210	.9271	.9329
SB	.2529	.2668	.2780	.2587	.2639
N = 2,000					
AMSE	1.0595	1.0692	1.0848	1.0601	1.0681
Variance	.7489	.7457	.7428	.7467	.7515
SB	.3107	.3234	.3420	.3134	.3167

across the methods in terms of AMSE, equivalents obtained from the integration method using the population distribution were more accurate than those using the posterior distribution or any of the summation methods. The equivalents given by the summation method with EAP were slightly better

Figure 4
MSE and SB at Each NC Score Level for Five Equating Methods With a SN Distribution



than those given by the integration method using the posterior distribution. The MSE and SB at each integer score point for the skewed population simulation with $N = 2{,}000$ are plotted in Figures 5a and 5b, respectively.

Figure 5a shows that the equating results for the summation method using the ML $\hat{\theta}$s had the largest MSE at the lower and upper ends of the score distribution. The equating results for the integration method using the posterior distribution of θ had slightly larger MSE around scores of 35 and 42. Figure 5b shows that the equating results for the summation method using the ML $\hat{\theta}$ had the largest SB at the lower and upper ends of the score distribution. The equating results for the summation method using the EAP and BME $\hat{\theta}$ had the smallest SB at the lower end of the score distribution. The equating results for the integration method using the posterior θ distribution had slightly larger SB around a score of 42.

Discussion

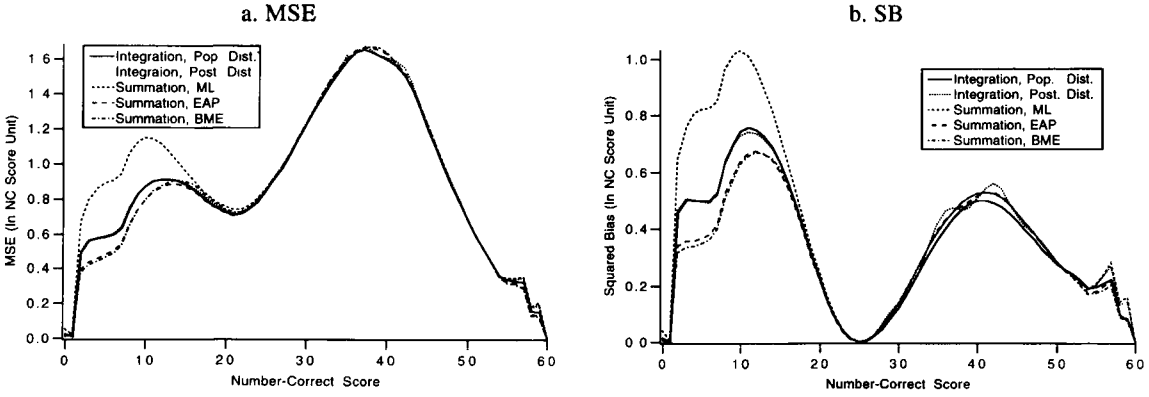
Errors in estimating item parameters and errors in estimating θ or the distribution of θ are the two major error sources in IRT observed-score equating. BILOG assumed θ to be normally distributed, but the data generated with the skewed population obviously failed to satisfy this assumption. Therefore, the posterior distribution of θ was not accurate, which resulted in less accurate equating results for the skewed simulation than for the SN simulation.

The results of the two simulations indicate that equatings using (1) the integration method with the actual population distributions or accurate estimates of the population distribution, and (2) the summation method with EAP estimates, were more accurate than the equatings using the summation method with the ML and BM estimates. One reason for this finding is that when the summation method was used, the errors involved in θ estimation were inherited in the equating process.

Lord (1982) pointed out that if the $\hat{\theta}$s estimated from one form contain larger or smaller errors of estimation than those from the other form, the two generated score distributions will not be properly comparable. The EAP estimate is defined as the mean of the posterior θ distribution given an observed response pattern. The average error in EAP estimates is smaller than that in the ML and BM estimates (Mislevy & Bock, 1990). This reasoning might explain why the errors in equating using EAP estimates were smaller than those using the ML and BM estimates. Because the integration method does not use $\hat{\theta}$ to generate score distributions, the error involved in the $\hat{\theta}$ estimation process will not affect the equating results. This reasoning might explain why the equatings using the integration method had smaller errors.

Kim & Nicewander (1993) compared ML, EAP,

Figure 5
MSE and SB at Each NC Score Level for Five Equating Methods With a Skewed Distribution



and BM θ estimators. They found that ML estimates yielded unacceptably low reliability coefficients relative to the other estimators they studied. In the present study, MSE and SB at NC score point levels indicated that the equating equivalents obtained from ML estimates had the largest MSEs and SBs at the two extremes of the score range (see Figures 4 and 5). Kim & Nicewander also reported that the θ estimators they studied were biased at both extremes of the score distributions for tests of moderate difficulty. The simulations in this study indicated that the equating equivalents obtained from the ML and BM $\hat{\theta}$s were biased at both extremes of the score distribution (see Figure 4b). Findings in Kim & Nicewander and this study suggest that IRT observed-score equating that uses ML $\hat{\theta}$s results in greater error than other θ estimation procedures.

Although it was not the focus of this study, IRT true-score equating is an often-used method in which scores are equated using the test response functions for the two forms (Lord, 1982). IRT true-score equating equates true scores not observed scores. However, true scores are not available in practice. Instead, ONCSs are converted using equating of true scores, although there is no theoretical reason for following this procedure. In addition, the true score relationship does not exist below the sum of the c parameters for tests that are fit using the three-parameter logistic model. One benefit of using IRT true-score equating is that the equating relationships are group invariant, assuming that the IRT model holds. In contrast to IRT true-score equating, IRT observed-score equating is intended to equate observed scores, and as such it is group dependent. As a practical matter, whether IRT observed-score or IRT true-score equating is more appropriate depends on the properties of the equating relationship that are considered most important for a particular equating. In the simulations presented here, the IRT observed-score equating relationship in the population (measured without error) was used as the criterion, because alternate ways to conduct IRT observed-score equating were being investigated. Harris & Crouse (1993) provided a discussion of various criteria that can be used in comparing equating methods and equating results.

Conclusions

The proposed IRT observed-score method using numerical integration to compute the observed-score distributions was found to be at least as accurate as the summation method using EAP θ estimates and more accurate than the summation method using ML and BM θ estimates. The integration method can be applied to equatings with the random groups design or the common-item nonequivalent-groups design. Also, the integration method can be implemented without estimating θ for individual examinees and it is computationally less intensive than the summation method.

References

American College Testing Program. (1989). *Preliminary technical manual for the enhanced ACT assessment.* Iowa City IA: Author.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Han, T. (1993). *Comparison of IRT observed-score equating with both IRT true-score and classical equipercentile equating* (Doctoral dissertation, Southern Illinois University, Carbondale, 1993). *Dissertation Abstracts International, 54/08,* 2997.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195–240.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58,* 587–599.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1–11.

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

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score

"equatings." *Applied Psychological Measurement, 8,* 453–461.

Mislevy, R. J. (1993). *Some formulas for use with Bayesian ability estimates* (Research Rep. RR-93-3). Princeton NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3* (2nd ed.). Mooresville IN: Scientific Software Inc.

Author's Address

Send requests for reprints or further information to Lingjia Zeng, American College Testing, 2201 N. Dodge St., P.O. Box 168, Iowa City IA 52243, U.S.A.