LSAC RESEARCH REPORT SERIES

■ Evaluating Equating Error in Observed-Score Equating

Wim J. van der Linden
University of Twente, Enschede, The Netherlands

■ Law School Admission Council
Computerized Testing Report 04-03
July 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

© 2005 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, Box 40, Newtown, PA 18940-0040.

LSAT and LSAC are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The statistical process called test equating is used to adjust for minor differences in difficulty between two forms of the same test and assures that test scores are comparable even though they have been derived from different test forms. The problem addressed in this research is how to define equating error if the criterion of successful equating is that one should not be able to distinguish between an equated score on a test form and a score on the test form to which it has been equated. We formulate two equivalent definitions of equating error based on this criterion. One definition focuses on the error in the equated scores; the other focuses on the error in the equating transformation.

These error definitions were used to evaluate traditional equipercentile equating, wherein a test form is equated to a previous test form by identifying the test scores on the two forms that share a common percentile, and two new conditional equating methods. These methods were evaluated using two test forms assembled from a previous item pool of the Law School Admission Test (LSAT). It was shown that, under a variety of conditions, the equipercentile method tends to result in a serious degree of bias and error, whereas the new methods are practically free of any error, except when the test to be equated has poorly discriminating items.

Abstract

Traditionally, error in equating observed scores on two versions of a test is defined as the difference between the transformations that equate the quantiles of their distributions in the sample and in the population of examinees. This definition underlies, for example, the well-known approximation to the standard error of equating by Lord (1982). But it is argued that if the goal of equating is to adjust the scores of examinees on one version of the test to make them indistinguishable from those on another, equating error should be defined as the degree to which the equated scores realize this goal. Two equivalent definitions of equating error based on this criterion are formulated. It is shown how these definitions allow us to estimate the bias and mean-squared error of any equating method if the response data fit an item-response theory model. An evaluation of the traditional equipercentile equating method and two new conditional methods for tests from a previous item pool of the Law School Admission Test (LSAT) shows that, under a variety of conditions, the equipercentile method tends to result in a serious bias and error, whereas the new methods are practically free of any error, except when the test to be equated has poorly discriminating items.

Introduction

The goal of observed-score equating is to adjust the observed number-correct scores on a new version of a test to make them indistinguishable from the scores the examinees would have had if they had taken an old version. Key questions in observed-score equating are therefore: When are scores of examinees on two versions of a test indistinguishable from each other; how do we evaluate a given score transformation; and what transformation realizes this goal best?

In the practice of observed-score equating, the benchmark among the methods of observed-score equating is generally considered to be equipercentile equating with a random-groups design. This method requires the distribution functions of the equated scores and the scores on the old version of the test to be identical for the population of examinees. To estimate the transformation, the two versions are administered to different random samples from the population, and the transformation is inferred from the sample distributions.

More formally, equipercentile equating can be defined as follows. Let Y be the observed score on the new version of a test for a random examinee from a population that has to be equated to the observed score X on an old version. These variables take possible values y and x, respectively. For convenience, we will also use Y and X to denote the versions of the tests themselves. The distribution functions of the two variables are denoted as $F_Y(y)$ and $F_X(x)$. For simplicity, throughout this paper we will assume that both functions are continuous and monotone. The transformation to be found equates the quantiles in the two distributions. If the distribution functions are monotonically increasing, the transformation is

$$x = \varphi(y) = F_X^{-1}(F_Y(y)), \tag{1}$$

where $\varphi(y)$ is the y score equated to the scale of X (Braun & Holland, 1982). The transformation in (1), which is monotonically increasing, maps the yth quantile in a distribution with distribution function $F_Y(.)$ to the same quantile in a distribution with distribution function $F_X(.)$.

Note for future reference that the transformation in (1) is from scale y to scale x. If it is applied to y, however, the random variable Y is transformed into a random variable $\varphi(Y)$ with a population distribution identical to the distribution of X.

Because the distribution functions of X and Y have to be estimated, the actual transformation used to equate the scores of X and Y is

$$x = \hat{\varphi}(y) = \hat{F}_X^{-1}(\hat{F}_Y(y)), \tag{2}$$

where $\hat{F}_Y(y)$ and $\hat{F}_X(x)$ are estimates of the two population distribution functions.

One option for estimating the population distribution functions in (1) is to use their sample equivalents. For a typical test length, however, this method involves the estimation of a large number of parameters. (To be precise, n parameters for the distribution function of a test of n items, namely the population proportions for each of the number-correct scores 0, 1, ..., n minus one because their sum is constrained to be equal to one.) Therefore, to get stable sample distribution functions, large samples are required, particularly if the tails of the distributions, where the proportions of examinees in the population tend to be smaller, matter. The sample size can be reduced somewhat by using a sensible smoothing technique, but the danger of exchanging inaccurate for biased estimators is inherent in any such technique. An alternative option is to assume that the population distribution functions belong to a parametric family of functions and use the sample to estimate its parameter values. If statistical techniques are available to test the assumption against response data, this option becomes efficient. This option is used in the empirical examples later in this paper.

The estimator in (2) implies the following definition of equating error:

$$\varepsilon(y) = \hat{\varphi}(y) - \varphi(y) = \hat{F}_X^{-1}(\hat{F}_Y(y)) - F_X^{-1}(F_Y(y)). \tag{3}$$

That is, all equating error is due to sampling fluctuations in the estimators $\hat{F}_Y(y)$ and $\hat{F}_X(x)$ relative to the population distributions $F_Y(y)$ and $F_X(y)$. If these estimators converge to their population equivalents, as they do if the sample distribution function is used as the estimator of the population function, equating error vanishes, and the equating becomes perfect. Observe that in (3) error is random across sampling of examinees from the population. Also, error is a *function* of y; for different values of y, equating error takes different values.

The definition in (3) is indeed the definition underlying the literature on observed-score equating error, which was put on statistical footing in Lord's (1982) seminal paper, in which he presented a large-sample approximation to the standard error of equating. The approximation is given by:

$$Var(\hat{\varphi}(y)) \approx \frac{F_Y(y)[1 - F_Y(y)]}{[f_X(x')]^2}\left(\frac{1}{N_X} + \frac{1}{N_Y}\right), \tag{4}$$

where

$$x' = F_X^{-1}(F_Y(y)),$$

$f_X(x)$ is the probability density of X, and N_X and N_Y are the sizes of the samples from the distributions of X and Y. Lord's derivation of (4), which assumes that the distribution functions in (2) are estimated by their sample equivalents, is based on a variance decomposition for the left-hand side of (4) to deal with the fact that (2) is a composite function with random error in each component. The result in (4) is for continuous test scores; for an approximation for discrete scores, see Lord (1982).

A similar definition of equating error is found in Kolen and Brennan (1995, p. 212), though these authors arrive at the definition from the opposite direction. Given any transformation $\varphi(.)$, they define its application to a sample of examinees on test Y as the equated score $\hat{\varphi}(y)$. They then go on and, in a step reminiscent of the definition of an examinee's true score in classical test theory, define the true equated score as the expected value of the equated score across sampling, $\mathcal{E}[\hat{\varphi}(y)]$. It follows that the equating error associated with y is equal to

$$\varepsilon(y) = \hat{\varphi}(y) - \mathcal{E}[\hat{\varphi}(y)] \tag{5}$$

for all values of y. This definition is identical to the one in (3), provided $\mathcal{E}[\hat{\varphi}(y)] = \varphi(y)$. That is, $\hat{\varphi}(y)$ is an unbiased estimate of $\varphi(y)$ for all values of y, which holds if the distribution functions in (1) are estimated by their sample equivalents. To estimate the small-sample standard error associated with (5), Kolen and Brennan recommend using a (parametric) bootstrap estimator.

Formulation of Problem

The definitions of equating error above are puzzling for the following reasons. First, if the goal of observed-score equating is to yield equated scores on the new version of the test that, for all examinees, are indistinguishable from the scores on the old version, the definition of error should be based on this goal. That is, error should be defined as a measure of the difference between the equated score and the score on the old version of an exam.

Second, the definitions of equating error in (3) and (5) are based on a necessary condition but not on a necessary and sufficient condition for successful equating. If there are no equating errors in the sense that the equated scores on the new version of the test and observed scores on the old version are indistinguishable for each examinee in a population, the population distributions of these two scores are identical. But the reverse does not hold: If for a population the two distributions are identical, it is not necessary that the equated scores be indistinguishable from the scores on the old version of the test for each examinee. As we will discuss in the next section, observed scores have error components and should be considered as random variables. Equating their marginal distributions does not mean that these variables have been equated.

Third, the approach underlying (5) seems to be circular. It accepts any function of y as an equating transformation and defines its expectation across sampling as the true equated score against which the sample result should be evaluated. This approach works fine in classical test theory where, for lack of an external criterion, we replace the observed score of an examinee by its expectation as the parameter of interest, but not in observed-score equating, where we do have an external criterion in the form of the observed score on the old version of the test. Also, (5) implies that equated scores can never be biased, even if they result from an obviously wrong transformation, such as $\varphi(y) = c$, where c is the same arbitrary number for all values of y. In fact, for this transformation the definition in (5) even implies error-free equating for all values of y!

The problem addressed in this research is how to define equating error if the criterion of successful equating is that one should not be able to distinguish between equated scores on a new version of the test and scores on the old version. We formulate two equivalent definitions of equating error based on this criterion. One definition focuses on the error in the equated scores; the other focuses on the error in the equating transformation that creates the error in the scores. Further, we show that if these definitions are embedded in the framework of item response theory (IRT), it becomes possible to estimate the bias and mean-squared error in any type of equating. Though the notion of a standard error of equating is omnipresent in the literature, the idea that equating can be biased has drawn far less attention. We illustrate the procedure for the equipercentile transformation in (2) and two alternative equating transformations that are introduced below, using tests that were systematically varied in some of their properties.

Defining Equating Error

We will now make more precise what we mean when we say that observed scores on two versions of a test are "indistinguishable." Let P be the population of examinees from which we sample and p be an arbitrary examinee in this population. A basic assumption in test theory is that population distributions of observed scores have a two-level structure: At the level of an individual examinee, the observed score contains measurement error and is random over replications. At the level of a population of examinees, the observed score is a random variable with a distribution that is the result of a marginalization of the individual scores over their systematic or true components.

Hence, in observed-score equating study, for each examinee p, we have two random variables X_p and Y_p representing their scores on the old and new version of the test. In fact, we even have a third random variable. The equating transformation used in the study, $\varphi(y)$, transforms the score Y_p into a random variable $\varphi(Y_p)$, which is the equated score for examinee p.

An obvious criterion for a successful equating of the score of p is the difference between the equated score $\varphi(Y_p)$ and the score of p on the old test form, X_p. Our definition of "indistinguishable scores" focuses on this difference. We consider equated scores on the new version of the test and scores on the old version as indistinguishable for a population of examinees P if

$$X_p \text{ and } \varphi(Y_p) \text{ are identically distributed for all } p \in P. \tag{6}$$

This definition was documented earlier as a criterion of equating by Lord (1980), who called it the equity criterion and claimed that no transformation satisfying this criterion exists unless the two test versions are parallel and equating is not needed. As will be explained later in this paper, however, this claim was based on an unnecessary assumption about the nature of the transformation.

The following two sections present definitions of equating error that follow directly from (6):

1. *Error in Equated Scores*

Let $F_{X_p}(x)$ denote the distribution function of X_p, and $F_{\varphi(Y_p)}(\varphi(y))$ denote the distribution function of the equated score $\varphi(Y_p)$. The requirement in (6) that the variables X_p and $\varphi(Y_p)$ be identically distributed is met if their distribution functions are identical. Therefore, the difference between these functions is an adequate measure of equating error. Because $x = \varphi(y)$, we use $F_{X_p}(x) = F_{X_p}(\varphi(y))$ and evaluate the difference on the scale of y. Our definition of error in an equated score is:

$$\varepsilon_{p1}(y) = F_{\varphi(Y_p)}(\varphi(y)) - F_{X_p}(\varphi(y)). \tag{7}$$

Note that, as before, this definition leads to an error *function* for observed equating. But, unlike the error function for equipercentile equating in (3), we now have a different function for each examinee. Attempts to reduce this error function may lead to loss of information or wrong conclusions. For example, if we would focus only on the value of this function at the score by p on the new version of the test, that is, at the realization $Y_p = y$, we are forced to draw a wrong conclusion. For a discussion of this point, as well as of another potentially tempting but wrong definition of equating error, see the Appendix.

2. *Error in Equating Transformations*

If the condition $\varepsilon_{p1}(y) = 0$ is imposed on (7) for all y, and the equality is solved for $\varphi(y)$, the solution is an equating transformation free of error.

Setting $\varepsilon_{p1}(y) = 0$ gives

$$F_{X_p}(x) - F_{\varphi(Y_p)}(\varphi(y)) = 0. \tag{8}$$

Moving $F_{\varphi(Y_p)}(\varphi(y))$ to the right side and taking the inverse of $F_{X_p}(.)$ gives

$$x = F^{-1}_{X_p} F_{\varphi(Y_p)}(\varphi(y)). \tag{9}$$

This expression shows x as a function of y, but cannot be used as an equating transformation because it contains the unknown function $\varphi(.)$. However, as $\varphi(.)$ is monotone,

$$\begin{aligned} F_{\varphi(Y_p)}(\varphi(y)) &= \Pr\{\varphi(Y_p) \le \varphi(y))\} \\ &= \Pr\{Y_p \le y\} \\ &= F_{Y_p}(y). \end{aligned} \tag{10}$$

Substituting the result into (9) gives

$$\mathrm{x} = \varphi^*_p(y) = F^{-1}_{X_p}(F_{Y_p}(y)) \tag{11}$$

as an error-free equating transformation. We will refer to this transformation as the *true* equating transformation for examinee p.

Observe that this transformation has the same form as the equipercentile transformation in (1). The only difference is the replacement by the population distributions functions $F_X(x)$ and $F_Y(y)$ by the functions of the scores of p on the two tests, $F_{X_p}(x)$ and $F_{X_p}(\varphi(y))$, respectively. This analogy should not come as a surprise. The equipercentile transformation, which more appropriately should be called the quantile or the Q-Q transformation (Wilks & Gnanadesikan, 1968), can be used to transform any (continuous and monotone) distribution function into any other.

The true equating transformation $\varphi^*_p(y)$ in (11) can be used to evaluate the error in any other transformation. Let $\varphi(y)$ be a proposed equating transformation. As an alternative to (7), error in $\varphi(y)$ can be defined as

$$\varepsilon_{p2}(y) = \varphi(y) - \varphi^*_p(y) \tag{12}$$

$$= \varphi(y) - F^{-1}_{X_p}(F_{Y_p}(y)). \tag{13}$$

The two definitions in (7) and (12) are equivalent: $\varepsilon_{p2}(y)$ shows the error in the equating transformation $\varphi(y)$, whereas $\varepsilon_{p1}(y)$ shows the mismatch between the distributions of the equated score $\varphi(Y_p)$ and the score on the old test X_p that is the result.

Both error definitions lead to a function of y for each examinee. They differ in their range, though: $\varepsilon_{p1}(y)$ takes values in [–1, 1], whereas $\varepsilon_{p2}(y)$ takes values on the scale of X. In the remainder of this paper, we will focus on $\varepsilon_{p2}(y)$, because it is attractive to interpret equating errors directly on the scale of X. A graphical representation of the relation between the two error functions is given in Figure 1.

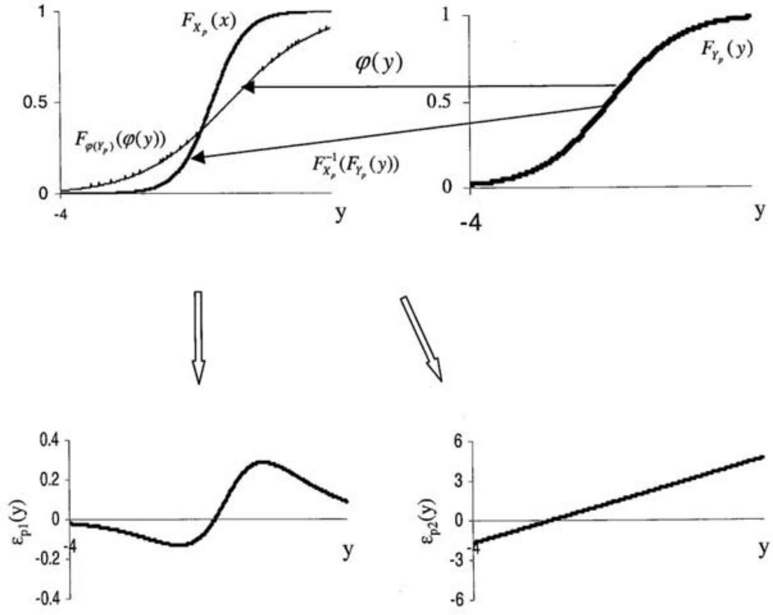


FIGURE 1. *Graphical representation of error in equated scores and in equating transformation*

Evaluating Equating Error

So far, the results have only been theoretical. In practice, the distributions of the observed scores X_p and Y_p are unknown, and the only data on p we will ever have is one realization of the scores $X_p = x$ or $Y_p = y$. Without any further assumptions, it is thus impossible to estimate equating error. We will now look into assumptions that allow us to do so.

IRT Observed-Score Equating

To illustrate the procedure, we suppose that the following three-parameter item response theory (IRT) model holds:

$$p_i(\theta) = \Pr(U_i = 1 \mid \theta) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i)]\}^{-1}, \tag{14}$$

where U_i is a binary variable for the response of the examinee to item i, $\theta \in (-\infty, \infty)$ represents the examinee's ability level, and $a_i \in [0, \infty)$, $b_i \in (-\infty, \infty)$, and $c_i \in [0,1]$ are the discriminating power, difficulty, and guessing parameter for item i (Lord, 1980).

The model is assumed to hold for X and Y simultaneously—that is, for the items in the two versions of the test calibrated together from a (possibly incomplete) set of response data. The requirement of simultaneous fit may seem strong. But observed-score equating can only be meaningful if the two versions of the test measure the same variable, and simultaneous fit is a powerful check on this condition. Also, if the model fits, there is no need to infer a functional relation between the examinee parameters for both tests; given θ, it is automatically clear what distribution of X is associated with what distribution of Y.

Under the model in (14), the distribution functions in (7) and (12) are replaced by the conditional functions of Y and X given θ, which will be denoted as $F_{Y|\theta}(y)$ and $F_{X|\theta}(x)$, respectively. These conditional distributions of Y and X are known to belong to the generalized binomial family, also known as the compound binomial family (e.g., Lord, 1980). This family does not have distribution functions that can be expressed in closed form, but one can easily calculate its members using a recursive procedure in Lord and Wingersky (1984). The procedure is based on the fact that the probabilities at $X = x$ are given by the coefficient of factor t^x in the generating function

$$\prod_{i=1}^{n}[q_i(\theta) + tp_i(\theta)], \tag{15}$$

with $q_i(\theta) = 1 - p_i(\theta)$.

Conditional Equating Methods

Using $F_{Y|\theta}(y)$ and $F_{X|\theta}(x)$, the definitions of equating error in (7) and (12) become

$$\varepsilon_1(y;\theta) = F_{\varphi(Y)|\theta}(\varphi(y)) - F_{X|\theta}(\varphi(y)) \tag{16}$$

and

$$\varepsilon_2(y;\theta) = (\varphi(y) - F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \tag{17}$$

respectively.

Observe that we no longer have different error functions of y for each examinee p but for each value of θ. As soon as the items have been calibrated from a sufficiently large calibration sample, we can calculate the two error functions for any value of θ from the parameter values. Unlike marginal IRT observed-score equating (Zeng & Kolen, 1995), no assumption whatsoever about the population distribution $g(\theta)$ is necessary. These functions are thus easily available to evaluate a new equating procedure or to choose the best procedure for an equating study from an available set of candidates.

The second error definition in (17) is based on the following set of true equating transformations

$$\varphi^*(y;\theta) = F^{-1}_{X|\theta}(F_{Y|\theta}(y)),\ \theta \in R. \tag{18}$$

This set was proposed directly for IRT observed-score equating by van der Linden (2000, Proposition 1), who proved that it meets all known criteria of perfect equating, namely (1) equity of equating for each examinee, (2) symmetry in X and Y, (3) population invariance, and (4) identical order of examinees on X and $\varphi(Y)$. For the first three criteria, see Harris and Crouse (1993) and Kolen and Brennan (1995).

The fact that (18) meets the criterion of equity seems to contradict Lord's (1980) theorem, which claims that, except for the trivial case of two parallel test versions, such transformations do not exist. Implicit in Lord's theorem, however, is the assumption that the transformation should be the same function for all examinees, whereas (18) implies different transformations for examinees with different ability levels θ.

The critical feature of the conditional equipercentile transformation in (18) relative to the marginal transformation in (1) is the use of the examinee's ability θ. The same idea of conditioning an equating procedure on a third variable was presented by Wright and Dorans (1993). A formal framework for doing so is presented by Liou, Cheng, and Li (2001). These authors motivated their approach by the wish to improve equating by accounting for relevant differences between examinees. Statistically, their idea amounts to the use of equating transformations conditional on values for background variables that describe relevant differences between examinees. In fact, the set of transformations in (18) takes this logic one important step further. Its transformations are conditional on the most relevant variable available; the ability of the examinees measured by the two versions of the test.

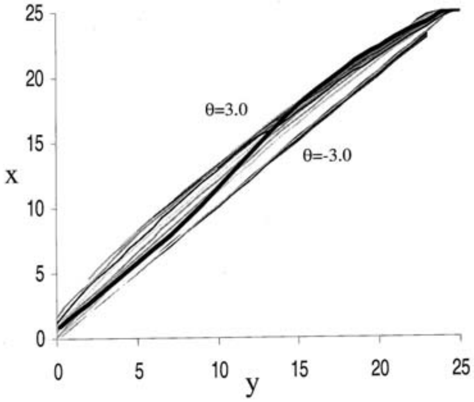


FIGURE 2. *Set of true conditional equating transformations conditional for two different versions of 25-item sections from the LSAT at θ values ranging systematically from $\theta = -3.0$ to $\theta = 3.0$, along with the marginal equipercentile transformation (bold curve)*

A graphical example of the true conditional transformations in (18) for a sample for θ values is given in Figure 2. The transformations are for two versions of a section of the Law School Admission Test (LSAT). The transformations were calculated using the Lord and Wingersky (1984) procedure discussed above. The only thing needed to calculate them was large-sample estimates of the item parameters. The figure also shows a marginal equipercentile transformation for the two sections. The differences between this transformation and the true transformations define its bias as a function of θ.

In actual testing, it is impossible to pick the exact transformation from the set in (18) for a given examinee because his/her true value of θ is not known. But we can use the information in the response vector of the examinee to approximate θ and make a best choice. Below, two approximations suggested by van der Linden (2000) are discussed. These approximations are expected to result in a smaller equating error than the estimated marginal equipercentile transformation in (2), which is to be applied as a common transformation to all examinees.

1. *Estimated Conditional Equating*

A simple approximation to (18) is to replace θ by a point estimate inferred from the examinee's response vector. In the empirical examples below, the mean of the posterior distribution (EAP estimator) was used to estimate θ. Let $\hat{\theta}$ denote the estimate. Upon substitution of $\hat{\theta}$ into (18), the following estimated conditional transformation is obtained:

$$\varphi(y;\hat{\theta}) = F^{-1}_{X|\hat{\theta}} F_{Y|\hat{\theta}}(y). \tag{19}$$

This transformation is based only on observable quantities. In an actual equating study, the following steps have to be taken to calculate an equated score for an examinee: (1) calculating the examinee's observed score on Y; (2) estimating his/her θ value from the response vector; (3) picking the true transformation at this estimate from (18); and (4) using this transformation to calculate the equated score associated with the examinee's observed number-correct score.

2. *Posterior Expected Conditional Equating*

The transformation in (19) uses the mean of the posterior distribution of θ but ignores the remaining uncertainty about θ in this distribution. From a Bayesian perspective, it seems fair to acknowledge this uncertainty and take the expectation of the set of true functions in (18) over the posterior distribution as an alternative to (19). If $f_{\Theta|u_1,\ldots,u_n}(\theta)$ denotes the density of the posterior distribution of θ for response vector $(u_1, \ldots, u_n)$, the equating transformation becomes

$$\varphi(y;u_1,\ldots,u_n) = \int F^{-1}_{X|\theta} F_{Y|\theta}(y) f_{\Theta|u_1,\ldots,u_n}(\theta)d\theta. \tag{20}$$

Estimating Equating Bias and RMSE

Traditionally, the interest in the statistical evaluation of equating error has been restricted to the standard error of equating. The fact that equating can be biased has never raised much concern. Even for the framework of marginal equating, the restriction is surprising. For example, linear equating transformations are low-parameter approximations to the marginal equipercentile transformation, and such approximations are bound to be biased, at least for certain portions of the score scale.

The fact that (18) contains the true equation transformation allows for a straightforward check on the bias in any other equating transformation. In fact, if this other equating would not involve any parameter estimation, all equating error would be systematic, and the functions in (16) and (17) could be interpreted directly as bias functions. Both in marginal equating and in the two conditional equating methods in this paper, we do need to estimate equating transformations, and this fact complicates the definition of their bias functions slightly. At the same time, we need a measure of their accuracy. In this paper we will use the root-mean-squared-error (RMSE) functions of the transformations as a measure of their accuracy.

In marginal equipercentile transformation, the transformation in (2) has to be estimated from the two population distributions of X and Y, and we have to evaluate the error function across sampling from these distributions. Let $\hat{\varphi}(y)$ denote the estimated transformation. Its bias function is

$$\text{Bias}(\hat{\varphi}(y); P) = \mathcal{E}_{X,Y;P}[\hat{\varphi}(y) - \varphi^{*}(y; \theta)], \tag{21}$$

where $\varphi^{*}(\text{y}; \theta)$ is the true equating transformation in (18) and $\varepsilon_{X,Y;P}$ denotes the expectation operation over X and Y with random sampling from population P. The root-mean-squared-error (RMSE) function is defined as:

$$\text{RMSE }(\hat{\varphi}(y); P) = \{\mathcal{E}_{X,Y;P}[\hat{\varphi}(y) - \varphi^{*}(y; \theta)]^{2}\}^{1/2}. \tag{22}$$

The standard error of equating can be calculated from these two quantities; it is the square root of the difference between the squared bias and MSE functions.

If a conditional equating transformation is evaluated, the two expressions change. For example, for the estimated conditional transformation $\varphi(y; \hat{\theta})$ in (19), they become

$$\text{Bias}(\varphi(y;\hat{\theta});\theta) = \mathcal{E}_{U_1,\dots,U_n|\theta}[\varphi(y;\hat{\theta}) - \varphi^{*}(y;\theta)]\,, \tag{23}$$

and

$$\text{RMSE}(\varphi(y;\hat{\theta});\theta) = \{\mathcal{E}_{U_1,\dots U_n|\theta}[\varphi(y;\hat{\theta}) - \varphi^{*}(y;\theta)]^{2}\}^{1/2}\ . \tag{24}$$

where $\mathcal{E}_{U_1,\dots,U_n|\theta}$ denotes the conditional expectation operation over random responses on Y given θ. Replacement of $\varphi(y; \theta)$ by $\varphi(y; u_1, \dots, u_n)$ in (20) gives the bias and RMSE functions for the posterior expected conditional equating transformation.

In (23) and (24), we ignored random variation due to estimation error in the item parameters. These expressions are thus for large-sample estimates of the item parameters. In the empirical study below, though, we show that the conditional equating methods are robust with respect to estimation error in the item parameters.

The expressions in (21) and (24) cannot be evaluated in closed form, but estimates with any desired precision can be obtained by Monte Carlo simulation of responses under the IRT model used and averaging the (squared) errors over responses. This method was used to evaluate the two conditional equating transformations against the marginal equipercentile transformation method in the following empirical study.

Empirical Examples

The bias and RMSE functions of the three transformations were evaluated for tests varying on the following factors:

(1) Length of X and Y;

(2) Difficulty of items in Y relative to those in X;

(3) Discriminating power of items in Y relative to those than in X;

(4) Size of error in parameter estimates of items in Y and X.

We expected the conditional equating transformations in this study to outperform the marginal transformation. As for the bias functions, the marginal equating is biased due to its dependence on the population distribution of θ. In addition, we expected the differences to increase as test length increased. For longer tests, the estimates of θ become more accurate, and so do the conditional transformations. On the other hand, if test length increases, the same happens to the number of parameters that define the distribution functions on which the marginal transformation in (2) is based, and less data per parameter is available to estimate them. Finally, we expected all three equating methods to be sensitive to the discriminating power of the items in Y. If the items become less discriminating, both the observed scores and estimates of θ for Y become less accurate, and the three methods are expected to perform worse than when discriminating power of items is higher. For completeness' sake, we also studied the effects of variation in the difficulty of the items in the two tests, and varied the errors in the parameter estimates of the items to evaluate the applicability of the conditional equating methods to tests with smaller samples of examinees.

Method

All versions of X and Y were derived from two blocks of 20 items selected from previous forms of the Law School Admission Test (LSAT). The 20-item, 40-item, and 60-item tests in the conditions with varying test length consisted of one, two, and three times the same block. This setup guaranteed conditions with homogeneous test lengthening and no confounding between test length and test composition.

The same 40-item versions of tests X and Y were used as a standard for comparison in all other conditions. The conditions with the more and less difficult versions of the items in Y were simulated by subtracting and adding .5 to the values of parameter b_i for the items in the standard test, respectively. The conditions with the more and less discriminating items in Y were simulated by multiplying their values for parameter a_i by .5 and 2.0, respectively. The resulting values are extremely small and large for commercial tests, respectively, but we exaggerated their size to make their effects clear. The condition with the smaller errors in the estimated value of the item parameters in X and Y was simulated by adding random numbers from the interval [–.15, .15] to the values of the items in the two standard tests for parameters a_i and b_i, and from [–.10, .10] to the values for parameter c_i. The condition with the larger errors was simulated by taking the intervals twice as wide. In both conditions, the values of the estimates of a_i and c_i were set equal to .10 and .00 if the result was lower than these values.

The conditional observed-score distributions of X and Y given θ were obtained for a fine grid of θ values using the Lord and Wingersky (1984) method for (15). The true conditional transformations in (18) at these values were calculated using the equipercentile method. To deal with the discreteness of the scores, we used the method of linear interpolation described in Kolen and Brennan (1995, chap. 2). For the two conditional methods, test administrations of Y were simulated for 5,000 examinees for each value $\theta = -2.00, -1.50, ..., 1.50, 2.00$. For each administration, the estimated and posterior expected conditional transformations in (19) and (20) were calculated, and the bias and RMSE functions of these transformations were evaluated by averaging their (squared) error over the administrations. The number of administrations for each value of θ was large enough to get stable estimates of these functions.

We calculated the marginal equipercentile transformation from the distributions of Y and Y, which were obtained by averaging the conditional distributions over the θ values. The transformation was calculated from these distributions, using the same method of linear interpolation as above. The marginal transformation was thus not the version estimated from the sample distribution functions in (2), but the population transformation in (1). Errors in this transformation were, therefore, entirely due to the differences between (1) and the true transformations in (18), not to estimation error of the distribution functions in (2). As a consequence, the comparison between the results for the marginal and conditional methods below is conservative in the sense that the marginal method was based on our knowledge of the true θ values of the examinees, whereas for the conditional methods estimates of θ were used.

If the conditional and marginal probability functions of X and Y yielded values smaller than .0001 for some x or y, these values were omitted from the study. For this reason, most of the figures do not display bias and RMSE functions over the entire range of y for all values of θ. This decision was motivated by the fact that, in a practical application, too few examinees would be available at these values, and the estimates of the transformations would have become unstable.

Results

The results for the different lengths of the test are displayed in Figure 3. The two conditional equating methods had negligibly small bias functions for all test lengths and values of θ. Their RMSE was somewhat larger, but always less than one score point on the scale of X for all values of y. The plots for the marginal equating method yielded curves for the bias functions that were generally ordered in θ, with the curves for the higher values of θ more to the left. Each curve showed a typical *wave*, which is the result of the difference in shape between the marginal transformation in (1) and the true conditional transformations in (18) (see Figure 2). As expected, these plots showed considerable bias at $n = 20$ for all values of θ, which further increased with the length of the test. The RMSE functions were of the same size as the bias functions, but lost their ordering property because of the operation of squaring involved in its calculation. The main conclusion from this part of the study is that the conditional methods are virtually unbiased but do have minor inaccuracy due to estimation of θ, whereas the marginal method is severely biased.

If we had used the estimated marginal transformation in (2) instead of the population transformation in (1), the marginal method would have shown larger values for its RMSE function due to estimation error.

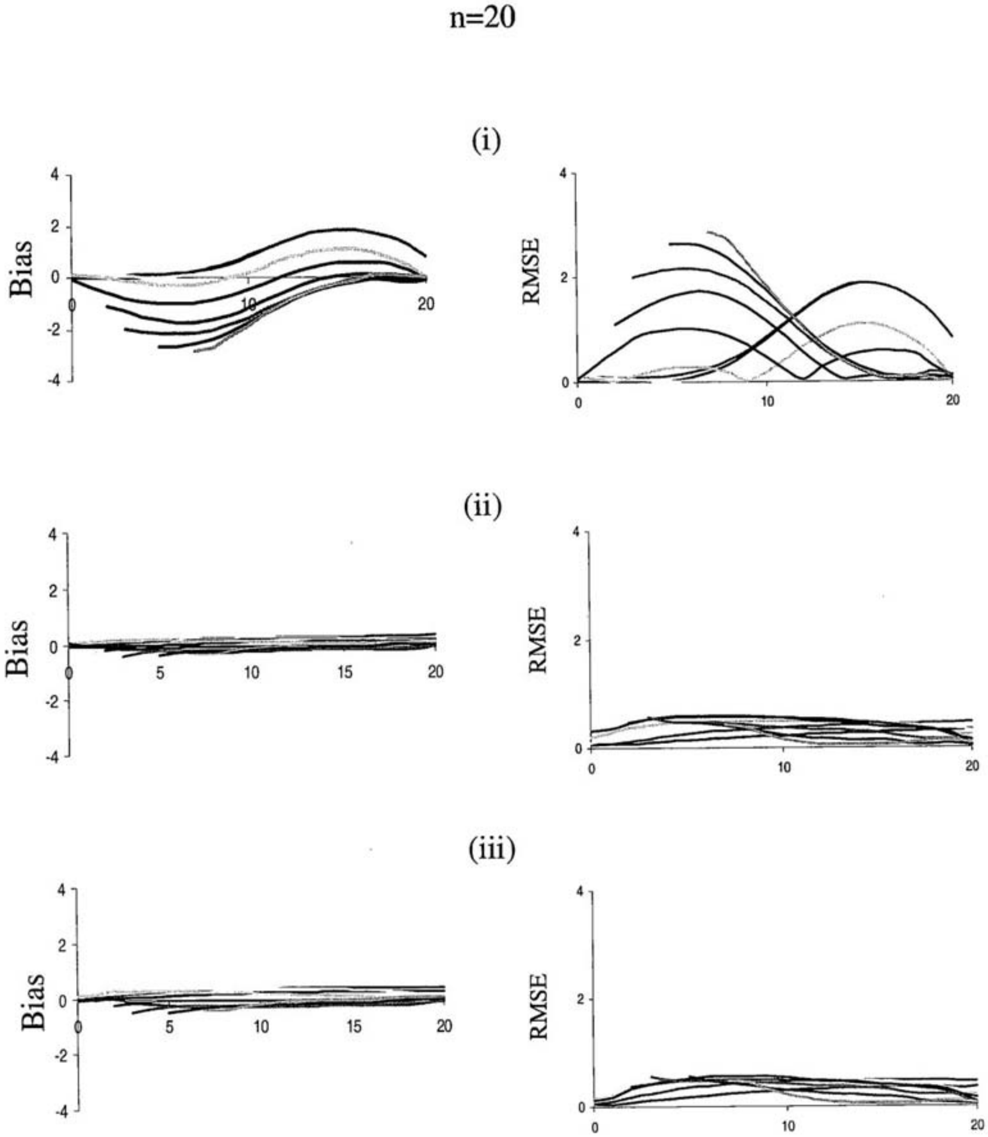


FIGURE 3. *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for test lengths n = 20 (Panel a), n = 40 (Panel b), and n = 60 (Panel c)*

(continued)

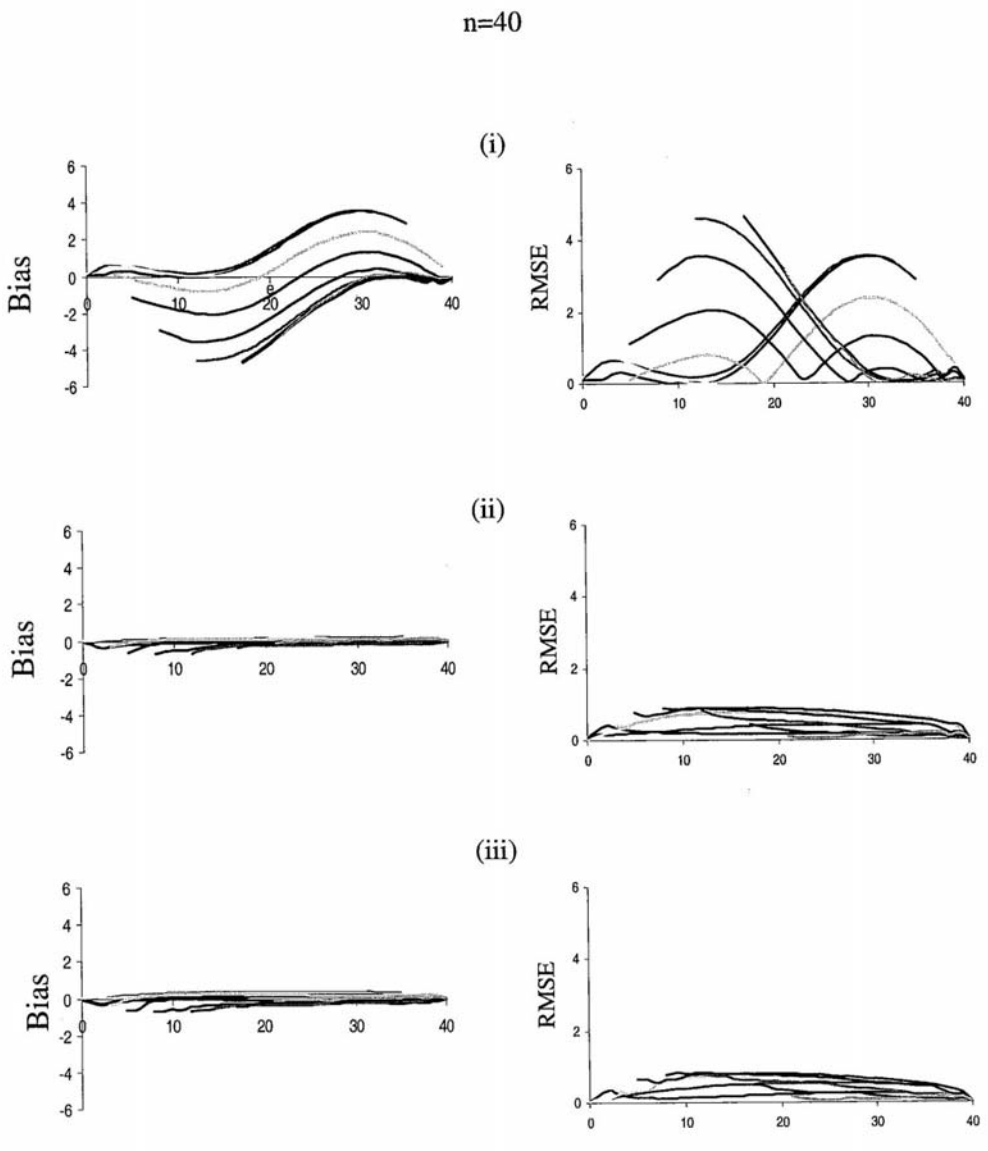


FIGURE 3 (continued). *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for test lengths n = 20 (Panel a), n = 40 (Panel b), and n = 60 (Panel c)*

(continued)

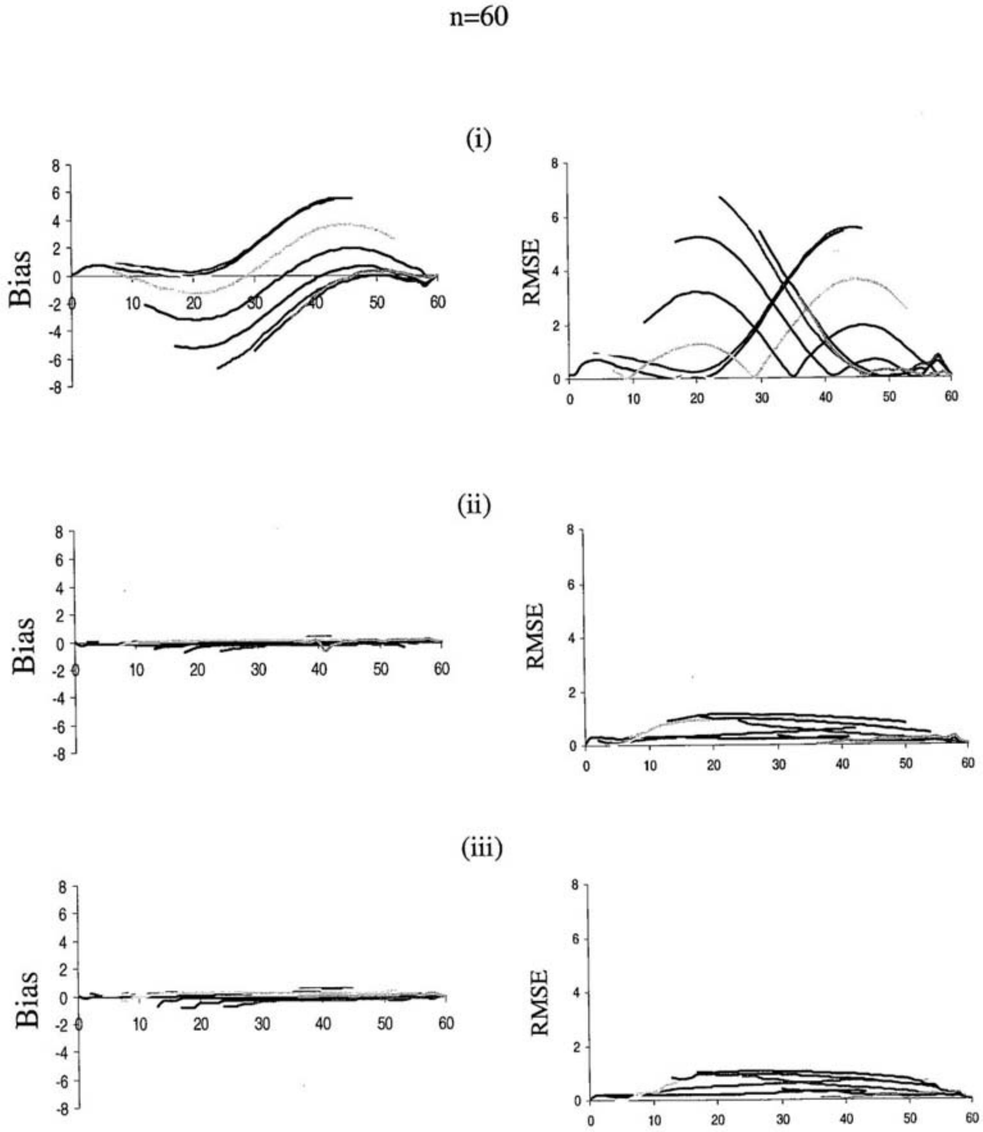


FIGURE 3 (continued). *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for test lengths $n = 20$ (Panel a), $n = 40$ (Panel b), and $n = 60$ (Panel c)*

Manipulating the differences in the values of the difficulty parameters between the two versions of the test did not result in any change in the bias and RMSE functions for the conditional transformations (see Figure 4). For the marginal transformation, the only result was a shift in the wave of the curves. This shift can be explained by the difference in shape between the marginal equating transformations for the cases with *X* less and more difficulty than *Y*. In the former case, the transformation is concave; in the latter, it is convex.

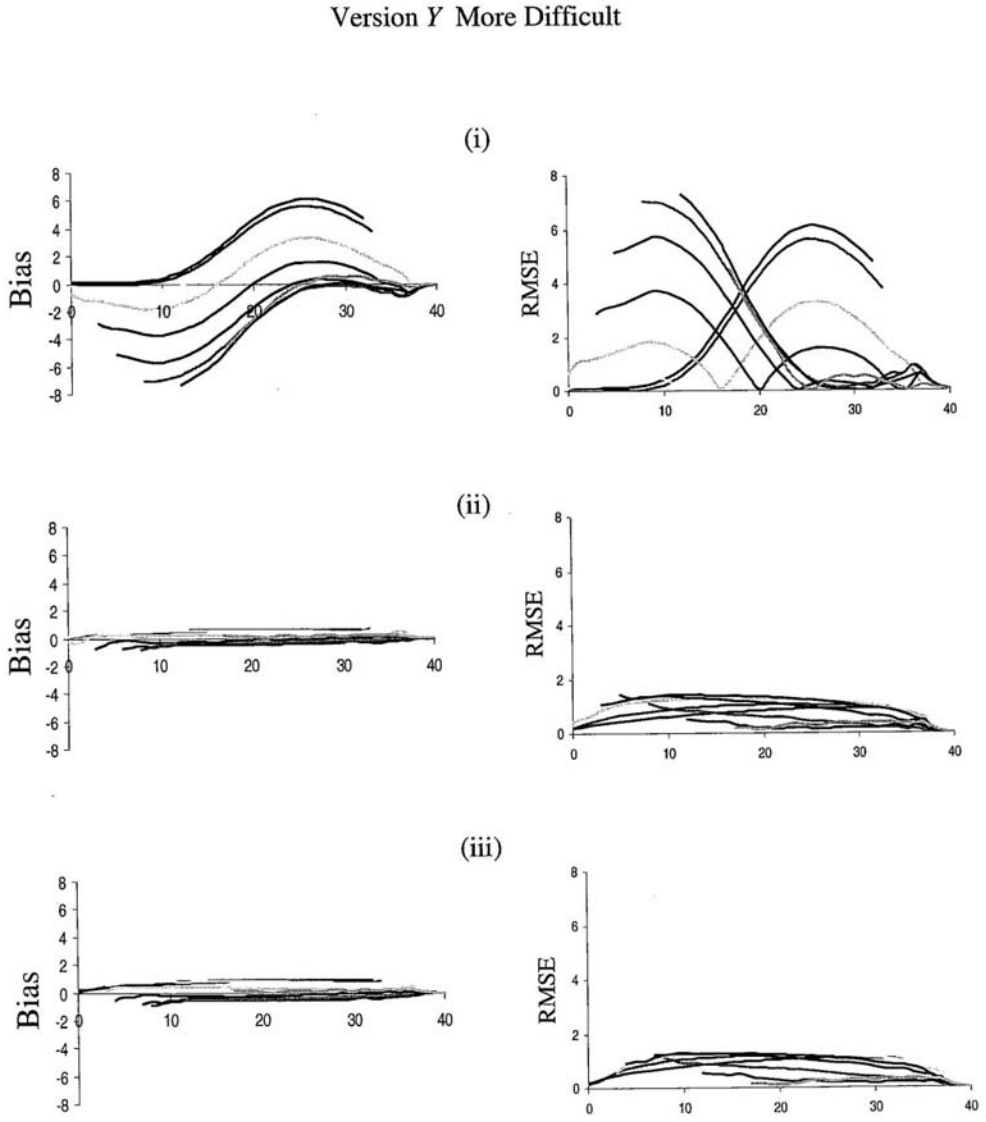


FIGURE 4. *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for a more difficult (Panel a) and easier (Panel b) new version of the test, Y (n = 40)*

(continued)

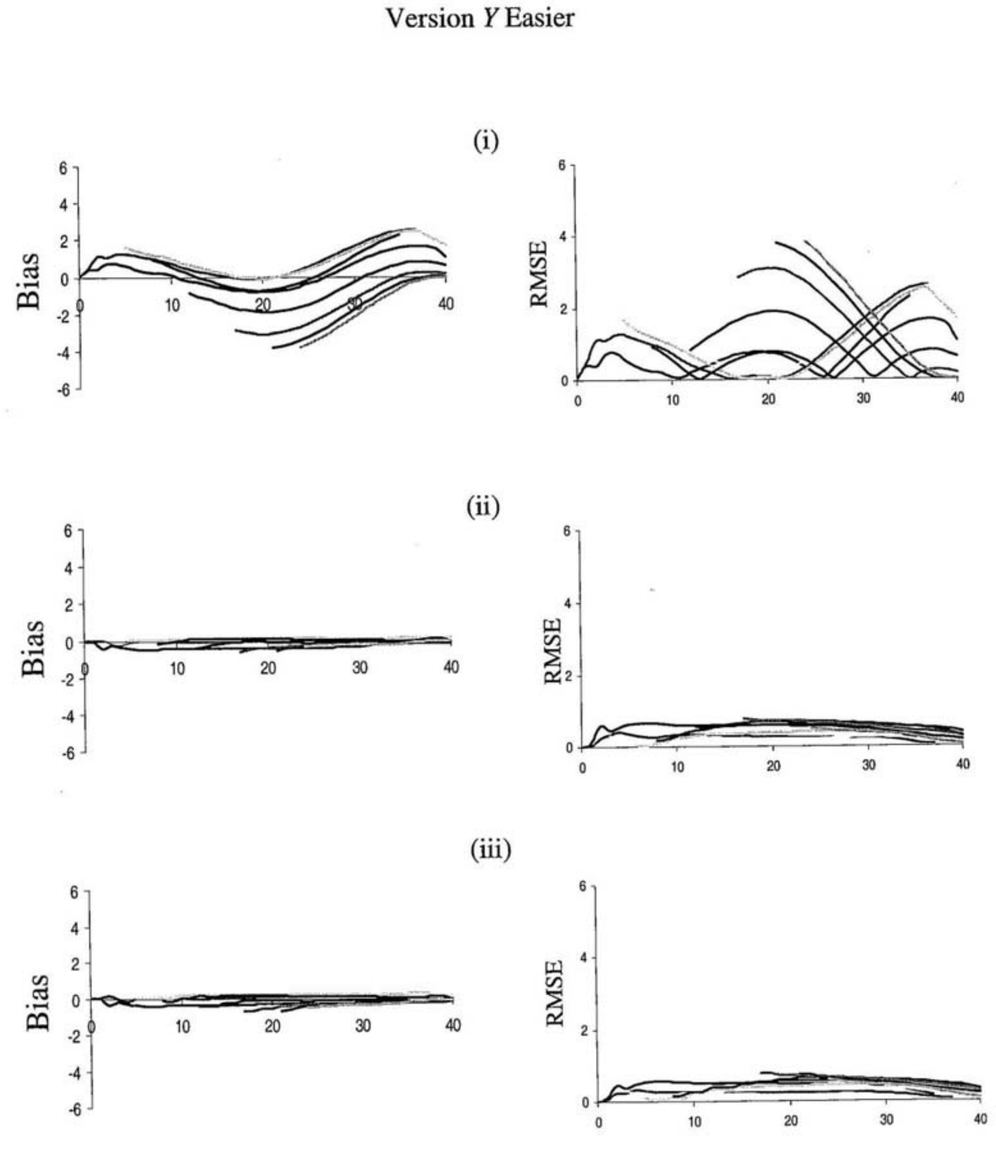


FIGURE 4 (continued). *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for a more difficult (Panel a) and easier (Panel b) new version of the test, Y (n = 40)*

Interesting results were obtained for the conditions in which the values of the item discrimination parameter for the new test, *Y*, were manipulated relative to those for the old test, *X*. Generally, lower values for this parameter means there are both wider observed number-correct distributions and fewer accurate estimators of *θ*. The first effect is visible in the bias and RMSE functions for the marginal equating method in Figure 5, which shows curves over larger ranges of *y* values than in the previous plot. The second effect explains the larger errors in the conditional equating transformations, which in this case were substantial, particularly for the lower values of *θ*. For larger values of the discrimination parameter, the opposite was observed: the marginal equating method still had large values for its bias and RMSE functions, but the curves were defined over much smaller ranges of values of *y*, whereas the conditional methods became virtually error-free. As noted earlier, the changes in the values of the discrimination parameter in these two conditions were not entirely realistic; they were made only to make the impact of the parameter on the equating transformations visible. In practice, it will seldom be possible to find commercial tests with parameter values that differ from the standard test in this study by a factor equal to .5 or 2.

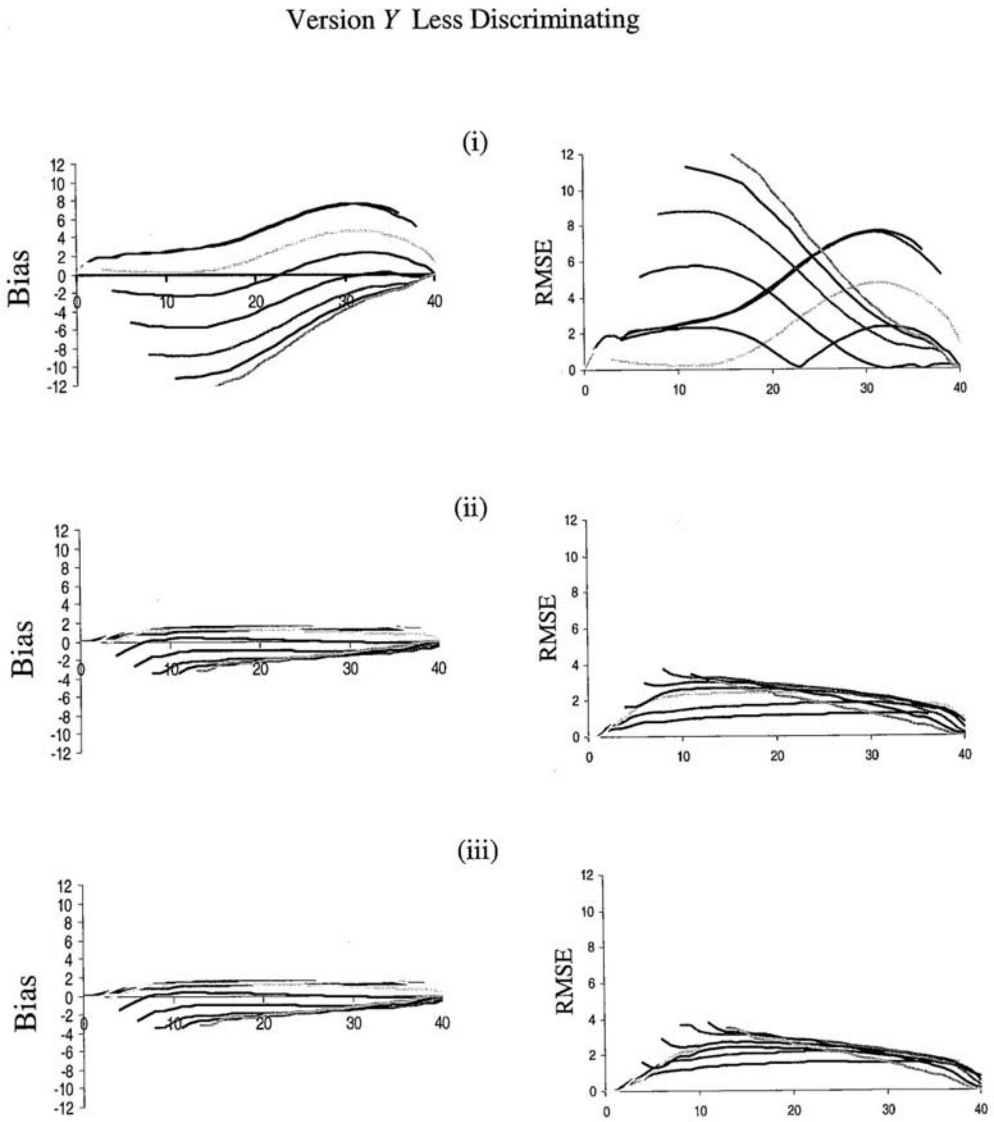


FIGURE 5. *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for a less discriminating (Panel a) and more discriminating (Panel b) new version of the test, Y (n = 40)*

(continued)

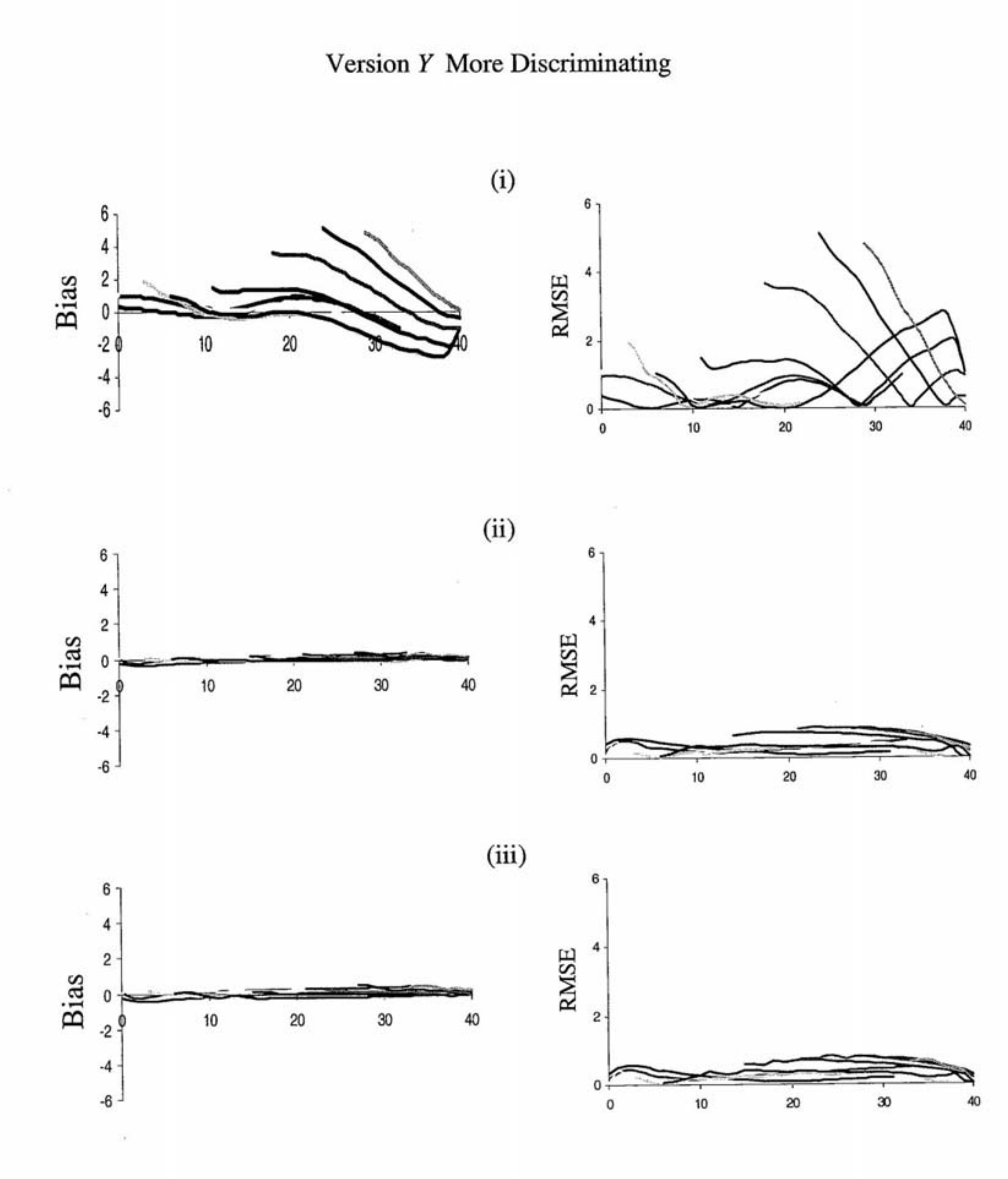
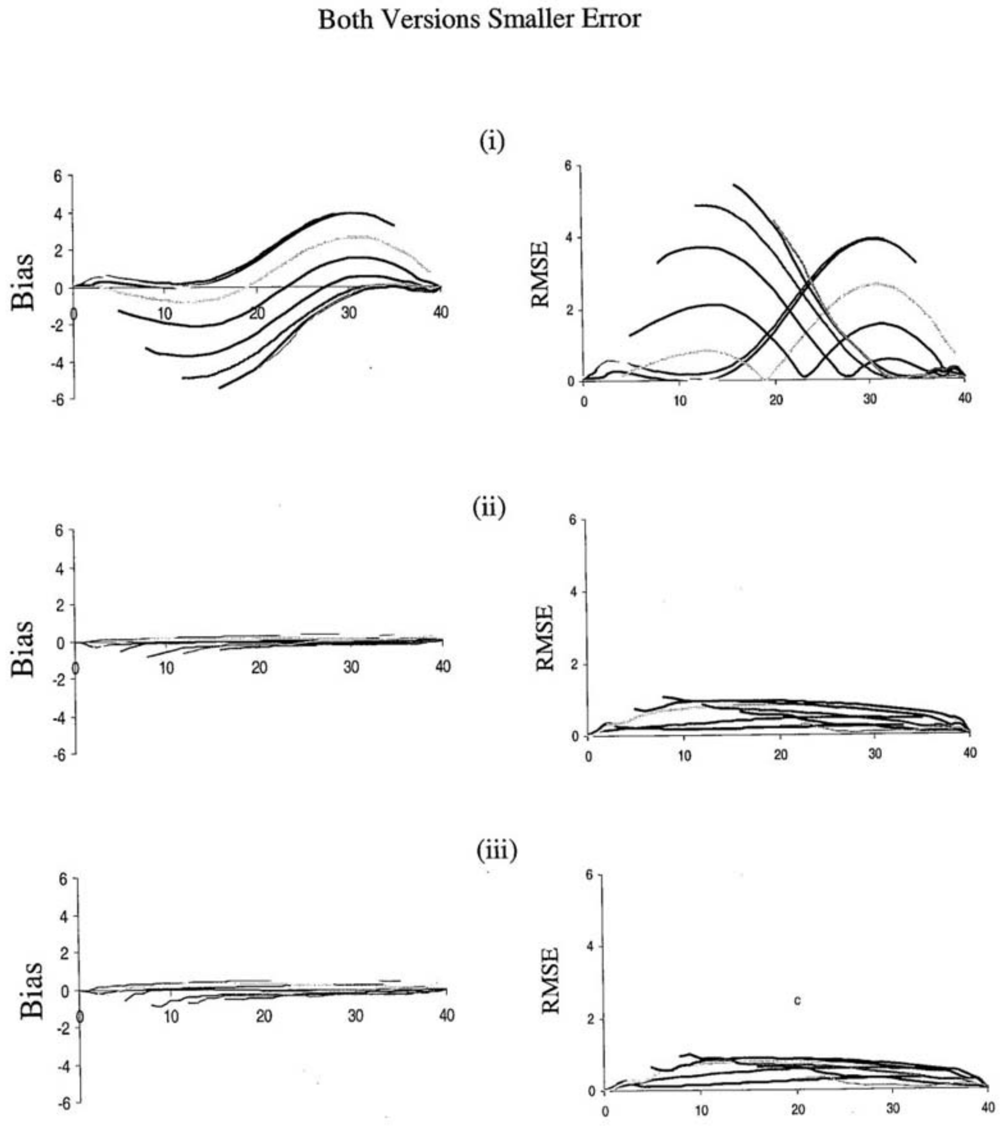


FIGURE 5 (continued). *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for a less discriminating (Panel a) and more discriminating (Panel b) new version of the test, Y (n = 40)*

The results in Figure 6 show that the presence of estimation error in the values of the parameters of the items hardly had any impact, even for the condition with the larger errors. All curves were basically the same as those for the 40-item standard test in Figure 3, which were obtained without any estimation error in the values for the item parameters. Apparently, the effects of estimation errors in the item parameters on the equating transformations have a tendency to average out.


Both Versions Smaller Error

(i)

(ii)

(iii)

FIGURE 6. *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for smaller (Panel a) and larger (Panel b) estimation errors in the values of the item parameters for the two versions of the test (n = 40)*

(continued)

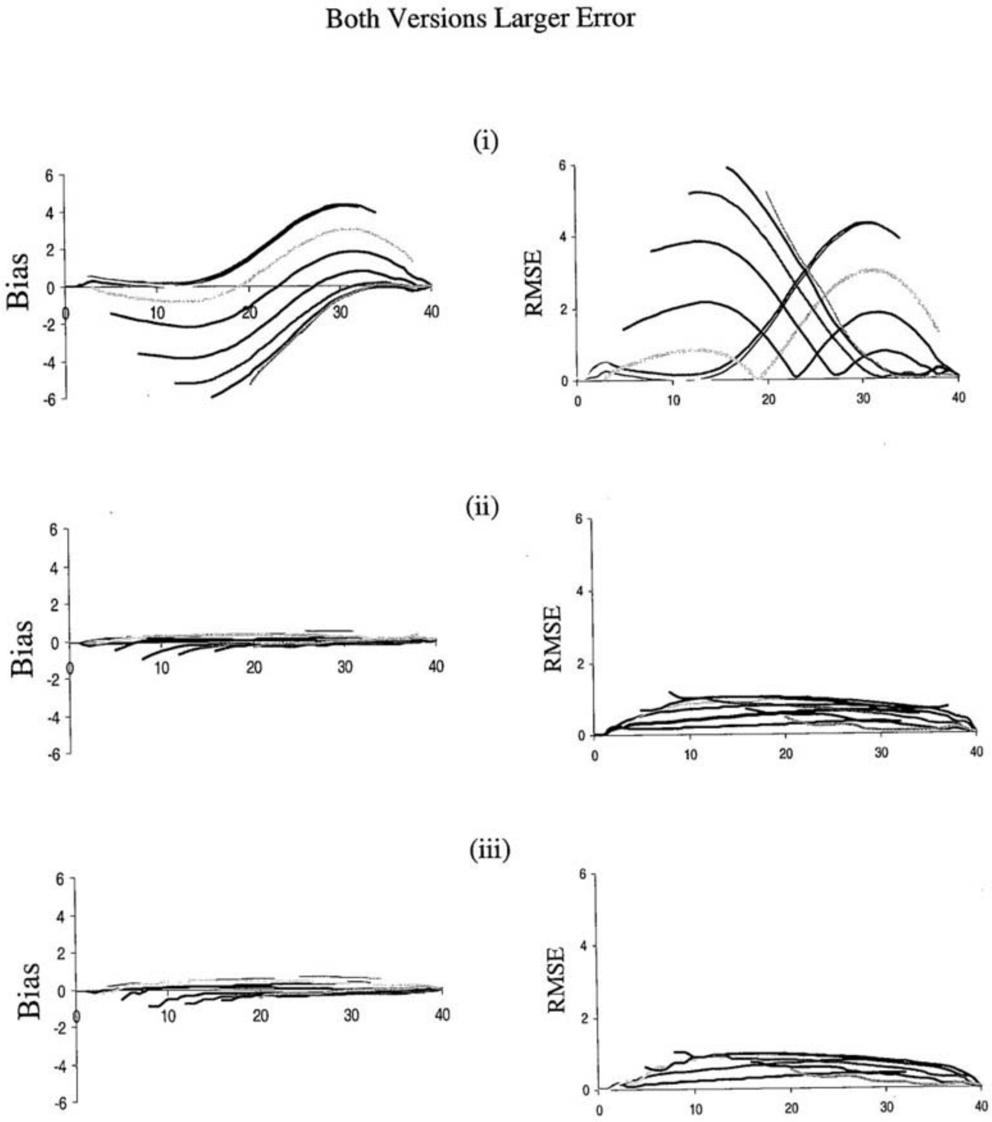


FIGURE 6 (continued). *Bias and RMSE functions for (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for smaller (Panel a) and larger (Panel b) estimation errors in the values of the item parameters for the two versions of the test (n = 40)*

Discussion

The research in this paper was motivated by the fact that the error definitions in the literature on observed-score equating allow for sampling errors in the estimates of population distribution functions for the two versions of the test only. Because the goal of equating is to transform the scores on the new version of a test to be indistinguishable from the scores on an old version, we argued that a definition of equating error based on the differences between the equated scores and the scores on the old version would be more natural. Embedding the definition in the framework of IRT, we derived the conditional equating transformation between X and Y, given θ as the true transformation for observed-score equating, and discussed two new methods of conditional equating that use estimates of θ to approximate the true transformation.

The results from an empirical study in which the bias and RMSE functions of the marginal equipercentile and the two conditional equating methods were evaluated under various conditions showed results that were generally in agreement with our expectations.

The conditional methods were virtually free of any bias and had only minor inaccuracy due to estimation of θ. The equipercentile method showed serious bias, which under some conditions was as large as 10–15% of the maximum score on the test to which they were equated. The only condition under which the conditional methods showed substantial error was for low values for the discrimination parameter for the items in the test to be equated.

A natural application for the conditional equating methods is in computerize adaptive testing (CAT), for example, when examinees are offered the choice between an adaptive and paper-and-pencil version of the same test, or when scores on an adaptive test are reported as number-correct scores on a reference test. This application of the methods, which would be an alternative to those studied in Lawrence and Feigenbaum (1997) and van der Linden (2001), is natural, because all items in the item pool are calibrated, and it is a relatively simple step for the computer algorithm to calculate the appropriate equating transformation along with the ability estimate on the adaptive version.

The reason observed-score equating should be done conditionally is because of random error in the observed test scores. Traditional equipercentile equating ignores the existence of this error; it assumes that the population distributions of X and Y contain fixed scores and all that needs to be done is to find the transformation that gives the two distributions the same shape. But examinees have different random error components in their scores, and these components entail different distributions of the scores on X and Y for different examinees. As a consequence, there is a need for a different equating transformation for each examinee, or, equivalently, a different transformation for each θ value in an IRT framework. Using a single transformation based on the marginal distributions of X and Y for all examinees leads to a necessary bias in their equated scores. The empirical study showed that the bias can be large under realistic conditions.

References

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.

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

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

Lawrence, I., & Feigenbaum, M. (1997). *Linking scores for computer-adaptive and paper-and-pencil administrations of the SAT* (Research Report No. 97-12). Princeton, NJ: Educational Testing Service.

Liou, M., Cheng, P. E., & Li, M.-Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25*, 197–207.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 7*, 165–174.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 452–461.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65*, 437–456.

van der Linden, W. J. (2001). Computerized adaptive testing with equated number-correct scoring. *Applied Psychological Measurement, 25,* 343–355.

Wilks, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55*, 1–17.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. 92-3). Princeton, NJ: Educational Testing Service.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*, 231–240.

Appendix

It may seem attractive to define equating error as the difference between the equated score associated with the examinee's *realized* score $Y_p = y$ on the new test and his/her score on the old test. But the score on the old test is a random variables X_p for which we do not have a realization. This definition of equating error would therefore amount to the following difference between the equating transformation $\varphi(.)$ evaluated at $Y_p = y$ and the random variable X_p:

$$\varepsilon_p \mid y = \varphi(y) - X_p. \tag{25}$$

As $\varphi(y)$ is a constant, (25) implies the same distribution of equating error for each possible value y. The standard deviation of this distribution is equal to $[\mathrm{Var}(X_p)]^{1/2}$. Thus, acceptance of (25) would imply a standard error of equating equal to p's standard error of measurement for the old test. Observe that this conclusion would hold for any equating transformation, even for one that would equate each score on Y to the same constant; that is, $\varphi(y) = c$. This implication shows that (25) cannot be used as a meaningful definition of equating error.

Another potentially tempting definition of equating error is to require not only that $\varphi(Y_p)$ and X_p be identically distributed, but also that they be *identical*. That is,

$$\varphi(Y_p) = X_p \text{ for all } p \in P. \tag{26}$$

If this requirement were to hold, equating error would have to be defined as

$$\varepsilon_p = \varphi(Y_p) - X_p. \tag{27}$$

But $\varphi(Y_p)$ and X_p are never identical. In fact, the well-known property of conditional (or local) independence of test scores even implies that they are independent. Because of this independence, the definition in (27) would lead to a standard error of equating equal to

$$[\mathrm{Var}(\varphi(Y_p)) + \mathrm{Var}(X_p)]^{1/2}, \text{ for all } p \in P. \tag{28}$$

This expression is minimal for the same transformation $\varphi(y) = c$. Its minimum is equal to $[\mathrm{Var}(X_p)]^{1/2}$, which is again the standard error of measurement for X_p. This implication shows that (16) cannot be used as a meaningful definition of equating error either.

Author Note

This paper was written while the author was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, California. The author is indebted to the Spencer Foundation for a grant awarded to the center to support his fellowship. The computational assistance of Wim M. M. Tielen and editorial comments of Kathleen Much are gratefully appreciated.