LSAC RESEARCH REPORT SERIES

■ Equating an Adaptive Test to a Linear Test

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

■ Law School Admission Council
Computerized Testing Report 04-01
December 2005

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.

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

A popular method of reporting scores on a computerized adaptive test (CAT) is to convert each test taker's CAT score to a number-correct score on a linear reference test; for example, a previous paper-and-pencil form or a special form assembled to the same content specifications as the CAT. This conversion is accomplished through a statistical process called test equating. Traditionally, two methods are used to carry out the equating that this score reporting method requires: (1) the equipercentile equating method, which determines the equated score by matching the percentiles for the CAT scores to those for the number-correct scores on the reference test, and (2) the test characteristic curve transformation equating method, which applies item response theory (IRT) to determine the reference test score that would correspond to each test taker's CAT score. (Note that IRT is a mathematical model that is commonly used to analyze test data.)

In this paper, it is argued that these two methods are necessarily biased because they use a single equating transformation for an entire population of test takers and therefore have to compromise between the observed-score distributions of individual test takers. Two new methods for the equating of an adaptive test to a linear test are presented, which allow for the differences between these individual observed-score distributions because they condition on the ability level of the individual test takers.

The four methods were evaluated empirically in a study with the difference between the distributions of the equated score and the actual observed score on the reference test as the success criterion. The four methods showed comparable degrees of precision but, as predicted, the two traditional methods were strongly biased while the two new methods were unbiased.

Abstract

Two new methods for the equating of an adaptive test to a linear test are presented. The methods are based on the conditional distributions of the observed scores on the two tests, given the examinee's ability. They are motivated by the fact that conditioning on the examinee's ability is necessary to allow for differences between observed-score distributions of examinees. The two methods were evaluated empirically against the traditional equipercentile method based on the marginal score distributions on the two tests and a method that uses the test characteristic function (TCF) of the linear test. The criterion in this study was the difference between the distribution of the equated score and the actual observed score on the linear test. The two conditional methods were unbiased and had mean-squared error in the equated scores comparable to the marginal equipercentile method and the TCF methods. The last two methods were strongly biased. It is argued that their bias is a consequence of the fact that they use a single equating transformation for an entire population of examinees and, therefore, have to compromise between the individual score distributions.

Introduction

In large-scale testing programs, test performances are generally reported by giving the examinees a sample of the test items along with the score calculated from their responses. If the test can be released after it has been administered, its items can be used for this purpose. Otherwise, an equivalent set of items has to be used.

In computerized adaptive testing (CAT), release of the test items after an examinee takes a test is impossible without immediate decrease in the quality of the item pool for the next examinee. To report scores, CAT programs therefore give their examinees a linear reference test (for example, a previous paper-and-pencil form or a special form assembled to the same content specifications as the CAT), and equate their CAT scores to a number-correct score on the reference test. Though a good reference test has to satisfy several potentially conflicting demands and its selection is not a trivial problem, this paper does not focus on this problem. Instead, it addresses the subsequent problem of identifying a method that can be used to equate CAT scores to number-correct scores on a chosen reference test. The same equating problem exists in testing programs that offer their examinees a choice between an adaptive and linear version of the same test.

In the CAT literature, three different methods for this equating problem have been investigated. The first method is equipercentile equating of the ability estimates on the CAT to the number-correct scores on the reference test. The method requires a separate empirical study prior to the operational stage of the test to estimate the population distributions of the ability estimates and number-correct scores on the two respective tests. The data in this study are usually collected following a randomly equivalent groups design or a design with nonequivalent groups but common items. The former has been used to equate the CAT version of the Armed Services Vocational Aptitude Battery (ASVAB) to its paper-and-pencil version (Segall, 1997); the latter, to equate the same two versions of the Scholastic Assessment Test (SAT) (Lawrence & Feigenbaum, 1997).

Suppose a randomly-equivalent-groups design is used to estimate the distributions. Let $F_{\hat{\Theta}}(\hat{\theta})$ be the distribution function of the ability estimates on the CAT for the population of examinees and $F_X(x)$ the function of the number-correct scores on the reference test, X. The transformation that has to be estimated equates the quantiles of the two distributions to each other. More formally, it is defined as

$$\hat{x} = e(\hat{\theta}) = F_X^{-1}(F_{\hat{\Theta}}(\hat{\theta})), \tag{1}$$

where $\hat{\theta}$ is the ability estimate equated to the scale of the number-correct score X and $e(.)$ is the transformation used to perform this equating. The transformation has the same general form as the traditional equipercentile transformation for number-correct scores on two linear tests (see, for instance, Braun & Holland, 1982); the only difference is the replacement of the distribution function of the number-correct score on the test from which the scores are equated by the function of the ability estimate, $\hat{\theta}$ on the CAT for the population of examinees.

Conducting an equating study to estimate (1) is usually expensive. To obtain accurate estimates of both distribution functions, large sample sizes are needed. If the sample sizes are too small, it may be possible to compensate somewhat by using an appropriate smoothing technique, but the choice of such a technique always has to strike a delicate balance between inaccurate and biased estimates of the distribution functions. In addition, each time the item pool, the population of examinees, or the CAT algorithm changes, a new equating study has to be conducted.

The second method is true-score equating using the test characteristic function (TCF) of the reference test. Suppose the items in the CAT pool and the reference test have been calibrated using the three parameter logistic response model

$$p_i(\theta) = \Pr(U_i = 1|\theta) = c_i + (1 - c_i)\left\{1 + \exp\left[-a_j(\theta - b_j)\right]\right\}^{-1}, \tag{2}$$

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 TCF for the reference test is defined as

$$\tau_X = \tau_X(\theta) = \sum_{i=1}^{n} p_i(\theta), \tag{3}$$

where τ_X is the classical test theory true number-correct score associated with X. In TCF equating, (3) is used to transform the examinee's ability estimate, $\hat{\theta}$, into an estimate of his/her true score on the reference test. The estimate is:

$$\hat{\tau}_X = \tau_X(\hat{\theta}) = \sum_{i=1}^{n} p_i(\hat{\theta}). \tag{4}$$

This method is less expensive than the previous one. The only requisite is previous calibration of the items in the reference test along with those in the CAT item pool. A subtle problem, however, is the difference between the estimated true score in (4) and the observed score on the reference test. Generally, these two scores are differently distributed for each examinee. If the two tests are not too short and the items in the reference test have been calibrated accurately, the differences between (4) and the observed score on the reference test are expected not to be large for most of the examinees, though.

The third approach is to impose constraints on the item selection in CAT that guarantee its observed number-correct scores to be automatically equated to the observed number-correct scores on the reference test. The constraints follow from a result in van der Linden and Luecht (1998). Let $i = 1, ..., n$ and $j = i, ..., n$ denote the items in the reference test and a CAT for an arbitrary examinee, respectively. These authors show that the observed score distributions on the two tests for an examinee with ability θ are identical if and only if

$$\sum_{i=1}^{n} p_i^r(\theta) = \sum_{j=1}^{n} p_j^r(\theta), \text{ for } r = 1, ..., n, \tag{5}$$

that is, if the two tests have equal sums of powers of the response probabilities of the order $r = 1, ..., n$ for the examinee. Because they also show that the importance of the higher-order powers vanishes quickly with increasing test length, typically for a real-life test, equality of the sums of powers of the first two or three orders is sufficient.

When applied to CAT, after each update of the ability estimate $\hat{\theta}$, the numbers on the left-hand side of (5), for $r = 1, 2, 3$, say, can be calculated from the response functions in the reference test. If the next item is selected, these numbers need to be imposed as constraint on the sums on the right-hand side of (5) at the current $\hat{\theta}$ for all items in the CAT. This seems hardly possible for traditional maximum-information CAT in which one item is selected from the pool at a time. But as the constraints in (5) are linear in the items, they can easily be imposed using the shadow test approach to CAT (van der Linden, 2000a; van der Linden & Reese, 1998). An empirical study (van der Linden, 2001) showed that this method yields excellent equating without any discernible loss of accuracy for the ability estimator for the test lengths typically used in CAT.

The focus of this paper is on equating methods based on score transformation. We study two new methods for equating CAT scores to number-correct scores on a reference test that can be used with any type of CAT. They are based on the same principle of equipercentile equating as in (1) but use the conditional score distributions given θ instead of the marginal distributions for the population of examinees to derive the equating transformation. A critical consequence of this choice is that, instead of a single uniform transformation for all examinees, we obtain a family of equating transformations with a different member for each value of θ. The statistical advantages of abandoning the traditional idea of a single transformation for an entire population of examinees are amply discussed in this paper. Also, from a more operational point of view, it is important that the conditional methods in this paper do not entail the necessity of any empirical equating study prior to the use of the test; neither do we have to re-estimate an equating transformation if the population changes, the item pool is replaced, or the CAT algorithm is modified. In this respect, these methods thus resemble the TCF method in (4). They differ from it, however, in that the latter is also based on a single transformation instead of a family of transformations and leads to the same type of statistical errors as the marginal equating method.

We will present results from CAT simulations to evaluate the bias and mean-squared error of all four equating methods discussed in this paper. As we will show later, our definitions of bias and mean-squared error are based on the notion of equating error as the difference between the distribution of the equated CAT score and the actual observed score on the reference tests for an examinee. We will also show that this notion is equivalent to the one of equating error as the difference between the equating transformation actually used and the true transformation developed for the case of equating observed scores on two linear tests presented in van der Linden (submitted).

Conditional Observed-Score Equating

We first motivate our choice of conditional equating over marginal equating for the case of two linear tests and then generalize the conditional methods to the case of equating an adaptive test to a linear test.

The conditional distributions of the number-correct scores on a test given θ belong to the generalized binomial family (also known as the compound binomial; see, e.g., Lord, 1980). We denote the conditional distribution functions for the scores X and Y given θ on two linear tests $F_{Y|\theta}(y)$ and $F_{X|\theta}(x)$, respectively, and assume that we equate from Y to X. These two functions can easily be calculated using a recursive procedure in Lord and Wingersky (1984). The procedure is based on the fact that the probabilities of $X = x$ are given by coefficients of the factor t^x in the expansion of the generating function

$$\prod_{i=1}^{n}\left[q_i(\theta) + tp_i(\theta)\right], \tag{6}$$

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

Using $F_{X|\theta}(x)$ and $F_{Y|\theta}(y)$, analogous to (1), the following family of *conditional* equating transformations can be defined:

$$x = e^*(\mathrm{y};\ \theta) = F^{-1}_{X|\theta}F_{Y|\theta}(y), -\infty < \theta < \infty. \tag{7}$$

If the items in the two tests have been calibrated, the response functions $p_i(\theta)$ in (6) are known. For a given value of θ we can then easily calculate the pair of distribution functions $F_{X|\theta}(x)$ and $F_{Y|\theta}(y)$, and, from them, the conditional transformation in (7). This point illustrates our earlier claim that, unlike the marginal equating transformation in (1), the calculation of a family of these transformations for a set of θ values does not involve the necessity of any empirical equating study with sampling of examinees for test X and Y. Also, because the family is defined conditionally on θ, it is insensitive to any change in the distribution of θ for the population of examinees.

Motivation

The family of transformation in (7) can be conceived of as the *true* equating transformation for test Y to X. If we would know the θ values of the examinees, use of these transformations would not involve any equating error. The family transforms the observed score on Y exactly into the score on X, and statistically it would be impossible to distinguish between the use of test Y with subsequent transformation of its scores and test X for any examinee (van der Linden, submitted). It should therefore not come as a surprise that this family meets each of the known criterion for perfect equating in the literature (van der Linden, 2000b, Proposition 1). First, as just indicated, it meets the criterion of identical error distributions for the equated scores and the scores equated to. This criterion is known as Lord's (1980; sect. 13.2) criterion of equity. Second, it meets the criterion of symmetry of the transformation in X and Y. If we exchange the role of X and Y in (7), we get the inverse of the transformation $e^*(y; \theta)$; that is, $e^*(x; \theta) = e^{*-1}(y; \theta)$. Third, as already indicated, conditional equating is population invariant; because the family in (7) is indexed by θ, it holds for any distribution over θ. For an introduction to these three criteria, see Harris and Crouse (1993) and Kolen and Brennan (1995). In addition, the family of transformations meets the criterion of identical (stochastic) order of the examinees by the equated score $e(Y)$ and the score on the test to which it is equated, X. This condition is equivalent to the condition of the two tests being unidimensional with a common dimension (van der Linden, 2000b).

The fact that the transformations in (7) take a different shape for each value of θ is necessary to meet these aforementioned criteria. Observed number-correct scores are random because of measurement error, and the scores of different examinees have different random components. If we use a single equating transformation for every examinee, we do not allow for these differences and are bound to create a bias in the equated scores. As we will show below, this bias can become extremely large for some examinees.

The choice between marginal and conditional equating has an instructive parallel in the one between marginal and conditional standard errors of measurement. Classical test theory was based on the assumption of a single standard error of measurement for the entire population of examinees. This assumption is obviously wrong. It creates the impression that all examinees are measured with uniform reliability, whereas we know, for instance, that observed scores are always bounded and examinees with more extreme abilities are measured less reliably than those with more moderate abilities. With the advent of item response theory, the standard error of measurement was redefined as the square root of the variance of the conditional distribution of X given θ, that is, $\mathrm{Var}(Y \mid \theta)$.

This expression shows intuitively better behavior and has been our choice of standard error ever since. Conditional equating is based on the same conditional distributions of observed scores as this standard error of measurement. It therefore automatically allows for differences between these distributions for different values of θ, whereas marginal equating assumes that they are identical.

Estimated Transformations

In operational testing, we do not know the true value of θ but estimate it from the examinee's response vector. This estimate can be used to make our best possible choice from the family of transformations in (7). The same practice is followed in the application of the TCF transformation in (4) and, continuing our parallel in the preceding section, if we report a standard error of a measurement for an ability estimate.

Depending on the type of estimation used, two different types of conditional equating are possible. If a point estimate of θ is inferred from the response vector, for example, the maximum likelihood estimate (MLE) or the Bayesian expected a posteriori estimate (EAP), our best choice from the family of equating transformations in (7) is:

$$e(y; \hat{\theta}) = F^{-1}_{X|\hat{\theta}} F_{Y|\hat{\theta}}(y). \tag{8}$$

If a fully Bayesian estimation procedure is used, it makes sense to allow for the remaining uncertainty about θ in the procedure and take the expectation of the transformations in (7) over the examinee's posterior. Let $f_{\Theta|u_1, \ldots, u_n}(\theta)$ denote the density of this posterior after the responses $U_1 = u_1, \ldots, U_n = u_n$ to the n items in the test. Our best choice of equating transformation then becomes

$$e(y; u_1, \ldots, u_n) = \int F^{-1}_{X|\theta} F_{Y|\theta}(y) f_{\Theta|u_1, \ldots, u_n}(\theta) d\theta. \tag{9}$$

We will refer to these two equation transformations as the *estimated conditional* and *posterior expected conditional transformation*, respectively.

Application to CAT

Statistically, CAT offers a natural environment for the application of conditional equating. For one thing, we always have an item pool fitting a response model such as the one in (2) with estimates of the item parameters good enough to treat them as their true values. Also, it is easy to extend the CAT algorithm with a procedure for calculating the conditional transformation in (8) or (9). The core of this procedure is Lord and Wingersky's (1984) algorithm for the distributions in (6), which involves only a few lines of computer code. If the transformation in (9) is used, it can be calculated using numerical integration over a well-chosen grid of θ values covered by the posterior distribution of the examinee.

It is interesting to note that the conditional equating transformations in (8)–(9) are not applied to $\hat{\theta}$ but to the number-correct score Y on the CAT. Hence, as an additional service to the examinees, conditional equating enables us to report to the examinees their *number-correct* score on the CAT along with the conversion table for the transformation that equates their score to a number-correct score on the reference test. The only thing the examinees need to understand is that because each of them takes a different set of items, they receive different conversion tables.

Evaluating Equating Transformations

The family of true equating transformations in (7) can be used as a benchmark to evaluate the errors in any transformation used in practice. A general treatment of the notion of error in observed-score equating is given in van der Linden (submitted), who derives two equivalent error functions. One error function represents the error in the equating transformation, the other in the equated score that results from the transformation. The formal definitions of these functions are repeated here.

Suppose we wish to evaluate an arbitrary transformation $\varphi(y)$ for the equating of an observed score on a test Y to a test X. The error in this transformation is the difference between $\varphi(y)$ and the true transformation $e^*(y;\theta)$ in (7). This definition leads to the following family of error functions:

$$\begin{aligned}\varepsilon_1(y;\theta) &= \varphi(y) - e^*(y;\theta)\\ &= \varphi(y) - F_{X|\theta}^{-1}F_{Y|\theta}(y),\ -\infty<\theta<\infty. \qquad (10)\end{aligned}$$

Observe that (10) consists of different *functions* of y for each θ.

The score of an examinee with ability level θ on test X has a known distribution with distribution function $F_{X\,|\,\theta}(x)$ defined by (6). On the other hand, if a transformation $\varphi(y)$ is used to equate the score of this examinee on test Y to test X, the result is an equated score $\varphi(Y)$ with distribution $F_{\varphi(Y)\,|\,\theta}(x)$. The difference between the two distributions functions defines a second family of error functions:

$$\varepsilon_2(x;\theta) = F_{\varphi(Y)\,|\,\theta}(x) - F_{X\,|\,\theta}(x),\ -\infty<\theta<\infty. \qquad (11)$$

These two families of error functions are equivalent in the sense that (l0) consists of the functions for the error in the equating *transformation* while (11) contains the functions for the error in the equated *score* that is the result of an application of the transformation.

When evaluating an equating transformation for two linear tests, the definition in (10) should be preferred. Its functions are defined on the scale of Y. A graph of these functions is thus easy to interpret. For example, it shows us for what regions the transformation stretches the scale of Y too much (i.e., the function has positive values) or too little (i.e., the function has negative values). For an empirical study with equating transformations for two linear tests based on this error definition, see van der Linden (submitted).

Choice of Error Function for CAT

Using the same error definition in an evaluation of the transformations for the equating of a CAT to a linear test studied in the current paper is impossible because the transformations are not on a common scale. This observation holds for two different reasons. First, in adaptive testing, different examinees take different sets of items and, therefore, the number-correct scale used in the conditional transformations in (8) and (9) is different for each examinee. In addition, the marginal and TCF transformations in (1) and (4) are not defined on the number-correct score scale but on the scale of the ability estimates $\hat{\theta}$.

For the equating of a CAT to a linear test in this paper, we use the alternative definition in (11). This family of error functions is not defined on the scale of Y but on the scale of X. These functions evaluate the impact of the transformations on the distribution of the equated scores, no matter what scores we equate from. They always take values in the interval [–1, 1]. The endpoints of the interval are approached when one of the distribution functions in (11) approaches its infimum and the other, its supremum.

We now discuss the shape of these error functions for the marginal and TCF transformations, as well as the two conditional transformations, and show how they can be calculated using Monte Carlo simulation.

Marginal and TCF Transformation

The distributions functions of the equated scores for the marginal equipercentile transformation $e(\hat{\theta})$ and the TCF transformation $\tau_X(\hat{\theta})$ for an examinee with ability θ are denoted as $F_{e(\hat{\Theta})|\theta}(x)$ and $F_{\tau_X(\hat{\Theta})|\theta}(x)$, respectively. From (11), the error functions for these equated scores are:

$$\varepsilon_2(x;\theta) = F_{e(\hat{\Theta})|\theta}(x) - F_{X|\theta}(x), -\infty < \theta < \infty \tag{12}$$

and

$$\varepsilon_2(x;\theta) = F_{\tau_X(\hat{\Theta})|\theta}(x) - F_{X|\theta}(x), -\infty < \theta < \infty. \tag{13}$$

For a given value of θ, the error functions can be calculated using the following procedure:

(1) Calculate the distribution function of the observed number-correct score on the reference test at θ, $F_{X|\theta}(x)$, using the Lord and Wingersky algorithm.

(2) Approximate the distribution function of the CAT estimates $\hat{\theta}$ at θ, $F_{\hat{\Theta}|\theta}(\hat{\theta})$, using a Monte Carlo method; that is, simulate CAT administrations for examinees at θ and record their estimates.

(3) Use the transformation for the marginal equipercentile method in (1) or the TCF method in (3) to transform the conditional distribution function $F_{\hat{\Theta}|\theta}(\hat{\theta})$ from Step 2 to the distribution function of the equated scores $F_{e(\hat{\Theta})|\theta}(x)$ or $F_{\tau_X(\hat{\Theta})|\theta}(x)$, respectively.

(4) Calculate the difference between $F_{e(\hat{\Theta})|\theta}(x)$ or $F_{\tau_X(\hat{\Theta})|\theta}(x)$ in Step 3 and $F_{X|\theta}(x)$ in Step 1.

The second step in this procedure can be executed with any required degree of precision by increasing the number of simulated CAT administrations.

Conditional Transformations

The error functions for the equated scores for the two conditional transformations in (8) and (9) are given by

$$\varepsilon_2(x;\theta) = F_{e(Y,\hat{\Theta})|\theta}(x) - F_{X|\theta}(x), -\infty < \theta < \infty \tag{14}$$

and

$$\varepsilon_2(x;\theta) = F_{e(Y;U_1,\ldots,U_n)|\theta}(x) - F_{X|\theta}(x), -\infty < \theta < \infty \tag{15}$$

respectively.

Observe that, unlike the marginal and TCF transformations, the error functions for the estimated conditional and posterior expected conditional transformations are random. They depend on the examinee's ability estimate $\hat{\Theta}$ and the response vector $(U_1, \ldots, U_n)$, respectively. For each CAT administration, we have a different realization of these functions.

For a realization of the error function for the estimated conditional transformation in (14), it holds that $F_{e(Y,\hat{\Theta})|\theta}(x) - F_{X|\theta=\hat{\theta}}(x)$. This conclusion follows immediately from (11) and the fact that $\varphi(Y) = X$. Hence, the realization is equal to the difference between the distribution of X at $\theta = \hat{\theta}$ and at the true value of θ. Or, formally,

$$F_{e(Y,\hat{\theta})|\theta}(x) - F_{X|\theta}(x) = F_{X|\theta=\hat{\theta}}(x) - F_{X|\theta}(x). \tag{16}$$

This conclusion leads to the following simple procedure for the calculation of a realization for an examinee at θ:

(1) Simulate a CAT administration for an examinee at θ and calculate the ability estimate $\hat{\theta}$.

(2) Calculate the distribution functions of the observed number-correct scores on the reference test at the ability estimate $\hat{\theta}$, $F_{X|\hat{\theta}}(x)$ and at the true ability θ, $F_{X|\theta}(x)$.

(3) Calculate the difference between the two distribution functions in Step 2. The result is the realization of the error function in (8).

The procedure for the calculation of a realization for the error function of the posterior expected conditional transformation consists of the following steps:

(1) Simulate a CAT administration for an examinee at θ, calculate the posterior distribution of θ, and record the items.

(2) Calculate the distribution functions of the observed number-correct score X on the reference test $F_{X|\theta}(x)$ and Y on the CAT in Step 1, $F_{Y|\theta}(y)$, for a grid of θ values using the Lord and Wingersky algorithm.

(3) Calculate the conditional transformations in (7) from the distribution functions in Step 2 for the same grid of θ values.

(4) Average the conditional transformations in Step 3 over the posterior distribution of θ in Step 1. The result is the posterior expected conditional transformation in (9) for the examinee in Step 2.

(5) Use the transformation in Step 4 to transform the conditional distribution function of number-correct score Y on the CAT in Step 1 at the true θ into the function of the equated score on X.

(6) Calculate the difference between the result in Step 5 and the distribution function for the number-correct score on X at θ, $F_{X|\theta}(x)$. The result is the realization of the error function in (15).

The distribution functions in the conditional transformations in (7) have no closed form. Consequently, the same holds for these transformations. Just like the marginal transformation in (1), which usually is based on empirical distribution functions with the same lack of closed form, we approximate them by finding corresponding quantiles in the two distributions and interpolating linearly to allow for the discreteness of the scores. For a description of this method, see Kolen and Brennan (1995, chap. 2).

The averaging in Step 4 leads to the necessity of finer interpolation than between the original discrete scores. If the impact of this interpolation is considered to be undesirable, we could approximate the error function in (15) by the difference between the posterior expectation of the distribution of the equated score and the distribution of X at the true value of θ. In doing so, we replace the result of an expected transformation with the expected result of a transformation, which leads to negligible error for the region of y values for which the transformation is close to linear. It is interesting to note that this approximation is in fact a Bayesian alternative to observed-score equating based on the conditional predictive probability function of X given ($U_1 = u_1, ..., U_n = u_n$) and θ (van der Linden, in preparation). The usual warning against the difference between regression functions and equating transformations in the classical texts on observed-score equating (e.g., Kolen & Brennan, 1995, p. 9) thus seems less urgent for the combination of conditional equating and Bayesian prediction.

Bias and MSE in Equating

Observed-score equating has a long tradition of evaluating equating methods only by their standard error of equating. This is correct as long as the method is known to be unbiased. In the empirical example below, we will show that this assumption is dangerous because traditional equating methods are strongly biased for some of the examinees. It is therefore more appropriate to evaluate equating methods by their bias and mean-squared error (MSE) or root-mean-squared error (RMSE) functions. As is well known, the standard error is related to these quantities in that its square is equal to the difference between the MSE and squared bias function.

Except for the numerical uncertainty in the Monte Carlo calculation of $F_{\hat{\Theta}|\theta}(\hat{\theta})$ in Step 2 above, the error functions for the marginal and TCF transformations in (12)–(13) are fixed quantities. As a consequence, these functions immediately show us the bias in these transformations, whereas their squares give their MSE function.

The two families of conditional transformations in (8)–(9) are random. Each CAT administration leads to a different transformation and, hence, involves a different realization of the error functions in (14)–(15). The bias and MSE functions for these transformations should therefore be defined as expected values over CAT administrations.

For the family of estimated conditional transformations, these expectations are defined as

$$\mathrm{Bias}(e(y;\hat{\theta});\theta) = \varepsilon_{U_1,\ldots,U_n|\theta}[F_{e(Y,\hat{\Theta})|\theta}(x) - F_{X|\theta}(x)]\,, \tag{17}$$

$$\mathrm{MSE}(\tau_X(\hat{\Theta});\theta) = \varepsilon_{U_1,\ldots,U_n|\theta}[F_{e(Y,\hat{\Theta})|\theta}(x) - F_{X|\theta}(x)]^2, \tag{18}$$

where $\varepsilon_{U_1,\ldots,U_n|\theta}$ denotes the expectation over replicated test administrations to examinees at θ. The definitions for the posterior expected conditional transformations are analogous.

Estimates of these bias and MSE functions are simply obtained by Monte Carlo estimation. They can be estimated as the mean and variance over a series of realizations of the error functions using the procedures outlined earlier. In the simulation study in the next section, we report the results from an empirical study in which these estimates were used to evaluate the statistical properties of the four equating methods investigated in this paper.

Empirical Study

An empirical evaluation of the following four methods of equating an adaptive test was conducted:

(1) The marginal equipercentile method in (1);

(2) The method based on the TCF in (4).

(3) The estimated conditional method in (8);

(4) The posterior expected conditional method in (9);

For each method, we estimated the bias and RMSE functions using CAT simulations for test lengths of n = 10, 20, 30, and 40. This range covers the actual test lengths used in most large-scale CAT programs. We chose the RMSE functions instead of the MSE functions because we wanted to be able to compare the bias and accuracy of the equatings on the same scale.

Item Pool and Tests

The simulations were conducted with an adaptive test from a previous pool of 753 items from the Law School Admission Test (LSAT). The pool was calibrated under the three-parameter logistic model in (2). The reference test was sampled randomly from an old form of the LSAT. Both the adaptive and the reference tests for the different test lengths were nested; the 20-item test always included the 10-item test, the 30-item test, the 20-item test, etc.

The ability estimator in the adaptive test was the expected a-posteriori (EAP) estimator with a uniform prior over [–4, 4]. The estimator was always initiated at $\theta = 0$. The items were selected using the maximum information criterion (Thissen & Mislevy, 1990).

Marginal Equipercentile Method

This method was applied for a population with $\theta \sim N(0, 1)$. To calculate the marginal equipercentile transformation, we approximated the population distribution function of the ability estimates, $\hat{\theta}$, by simulating 100,000 CAT administration for examinees randomly sampled from $N(0, 1)$. This large number was chosen to avoid sampling error in the equating transformation. The population distribution function of the number-correct scores on the reference test was calculated by averaging the conditional distributions for the examinees sampled from $N(0, 1)$. The marginal equipercentile transformation was calculated using the method of linear interpolation described in Kolen and Brennan (1995, chap. 2).

We evaluated this transformation at $\theta = -2.00, -1.50, \ldots, 2.00$. At each of the θ values, 20,000 CAT administrations were simulated to estimate the distribution function of $\hat{\theta}$ given the θ value.

The error functions in (12) at these values were calculated following the procedure outlined above. As already noted, the bias function is equal to this error function, whereas the RMSE function is equal to its absolute value.

TCF Method

The TCF transformation in (3) was calculated from the response functions of the items in the reference test.

This transformation was evaluated at the same values $\theta = -2.0, -1.5, \ldots, 2.0$ and using the same procedure as for the evaluation of the marginal transformation in the preceding section.

Conditional Methods

The two conditional methods were evaluated at the same θ values. At each value, we simulated 1,000 CAT administrations to estimate the bias and RMSE functions for the equated scores produced by these methods.

The realizations of the error functions for the simulated CAT administrations were calculated following the procedure outlined above. The posterior expected conditional transformations were calculated using numerical integration over a grid of θ values about the mean of the posterior distribution with stepsize .01.

As estimates of the bias and RMSE functions, the mean and standard deviation of the 1,000 realizations were used.

Results

Before presenting the estimates of the bias and RMSE functions, we discuss the plots of the distributions functions for the equated scores in Figure 1. These plots are for the adaptive tests with a length of $n = 20$. (The typical shapes of these plots help us to explain the behavior of the bias and RMSE functions in Figures 3–6.)

The left-hand side plots in Figure 1 show the distribution functions of the scores on the CAT after they have been equated to the linear test by the marginal and TCF method for $\theta = -2.0, -1.0, .0, 1.0$, and 2.0. (In this section, for brevity, we omit all plots for the remaining θ values because they invariably confirm the trends in the plots discussed in this section.) The right-hand side plots show the average distribution functions of the equated scores calculated over the CAT replications for the two conditional methods. Both plots also show the true distribution function for an examinee at these θ values, i.e., $F_{X|\theta}(x)$. This distribution function is our point of reference: the observed score of an examinee would have followed this function if he/she had taken the linear test directly. The difference between the distribution function of an equated score and the true distribution function is equal to the bias function of the equating method. The MSE function is the square of this difference for the marginal and TCF transformation, and the average square over CAT replications for the conditional transformations.

All distribution functions move to the right-hand side of the x scale with an increase in the values of θ. This is as expected because under the 3PL model in (2), more-able examinees have an observed-score distribution that is stochastically larger (that is, located more to the right) than less-able students. The largest bias was obtained for the marginal and TCF methods. Particularly, the TCF method had an extremely large bias at the lower tail of the distribution. The reason for this is the lower asymptote in the equating transformation in (3) due to the presence of the guessing parameter in the 3PL model in (2). The transformations for $n = 20$ found in this study are displayed in Figure 2. The major difference between them is the upward curve in the lower part of the TCF transformation due to this asymptote. The distribution functions for the two conditional methods show an unmistakable tendency to an observed-score distribution flatter than the true distribution. This tendency is a consequence of their dependence on estimation of θ. The tendency is slightly stronger for larger values of θ. This result is a consequence of the fact that the item pool is known to be on the difficult side for the θ values used in this study.

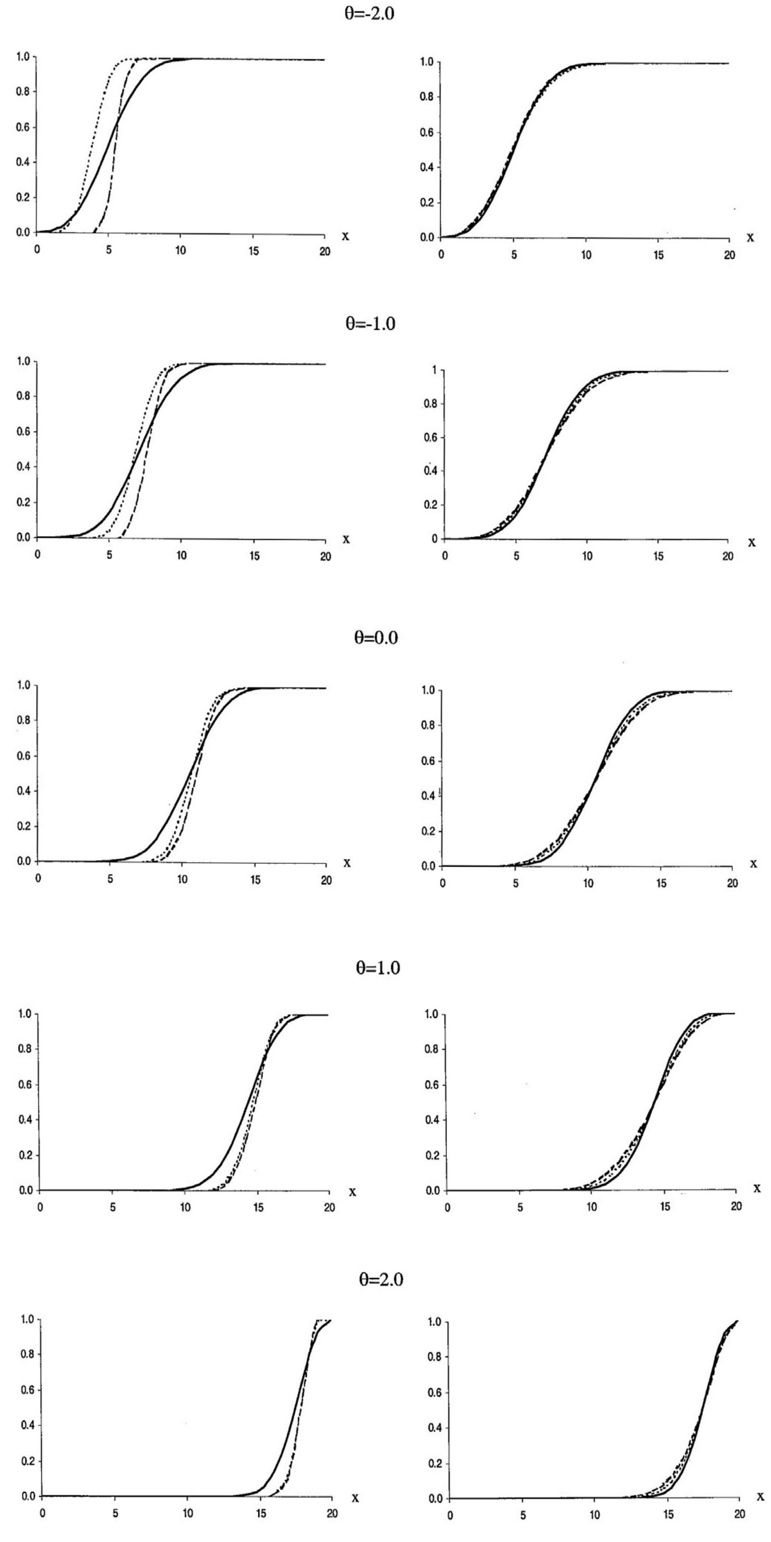

FIGURE 1. *Distribution functions of equated scores for the four equating methods in this paper for θ = –2.0, –1.5, ..., 2.0 and n = 20. Left-hand panel shows distribution functions for marginal method (dotted line) and TCF method (dashed line). Right-hand side panel shows average distribution functions for estimated conditional method (dotted line) and posterior expected conditional method (dashed line). Bold solid line in both panels is the distribution function for the examinee's actual observed score on the linear test*

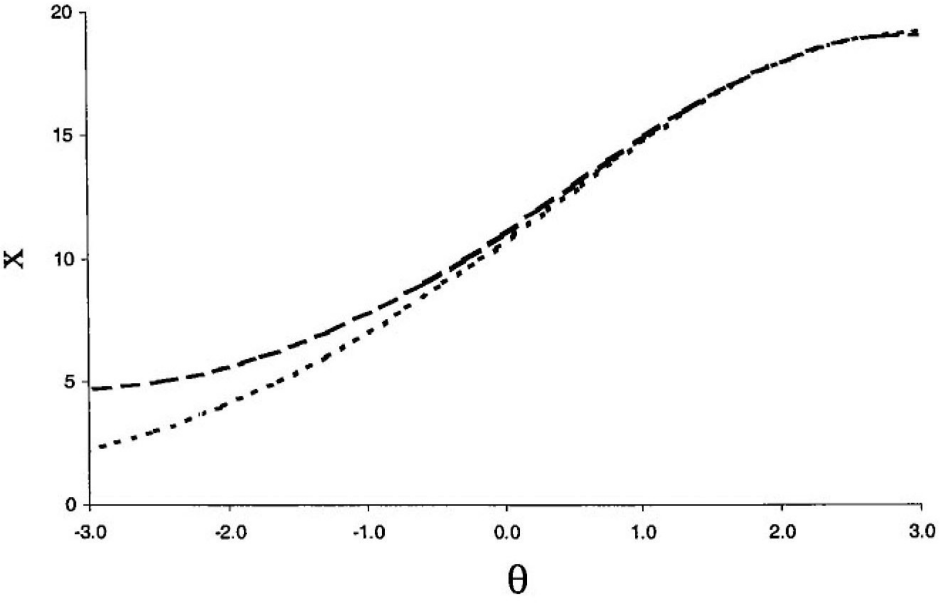

FIGURE 2. *Equating transformations for the marginal method (dotted line) and the TCF method (dashed line) for n = 20*

The bias and RMSE functions for test lengths n = 10, 20, 30, and 40 are shown in Figures 3 through 6. The functions for the marginal and TCF method and for the two conditional methods behave as two distinct groups. Several of the bias functions for the first two methods show a characteristic wave that is explained by the difference between their distribution functions and the function for the true distribution in Figure 1. The fact that the bias function for the TCF method begins with a negative peak follows from the lower asymptote that its transformation exhibits (see Figure 2). Hardly any bias is found for the two conditional equating methods, particularly for the lower θ values. For the larger θ values, the bias functions show a small wave, which is a consequence of the tendency to the flatter observed-score distribution for the CAT administrations at these values (indicated in the preceding paragraph).

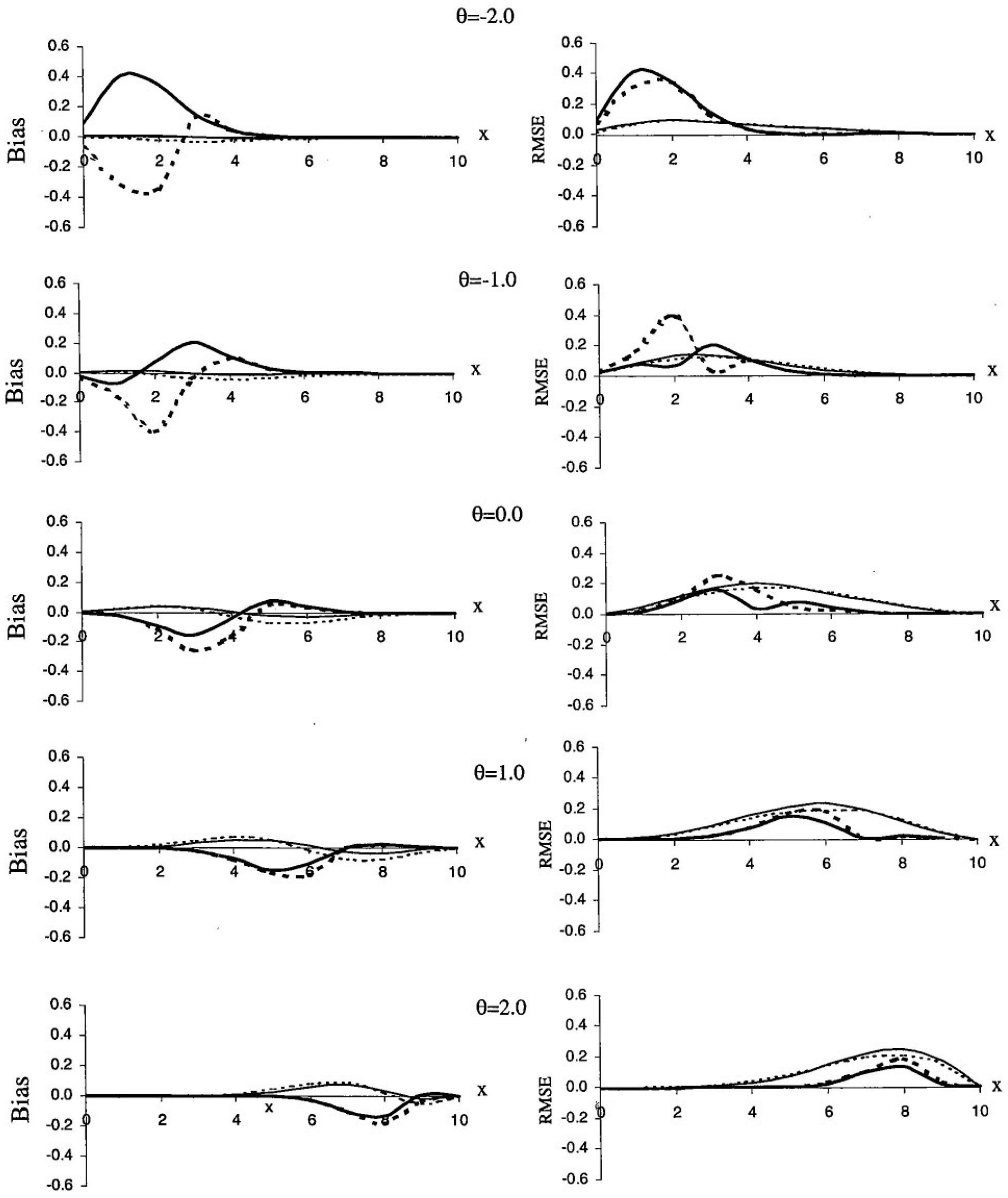

FIGURE 3. *Bias and RMSE functions of marginal equating method (bold solid line),TCF method (bold dotted line), estimated conditional method (thin solid line), and posterior expected conditional method (thin dotted line) for θ = –2.0, –1.5, ..., 2.0 and n = 10*

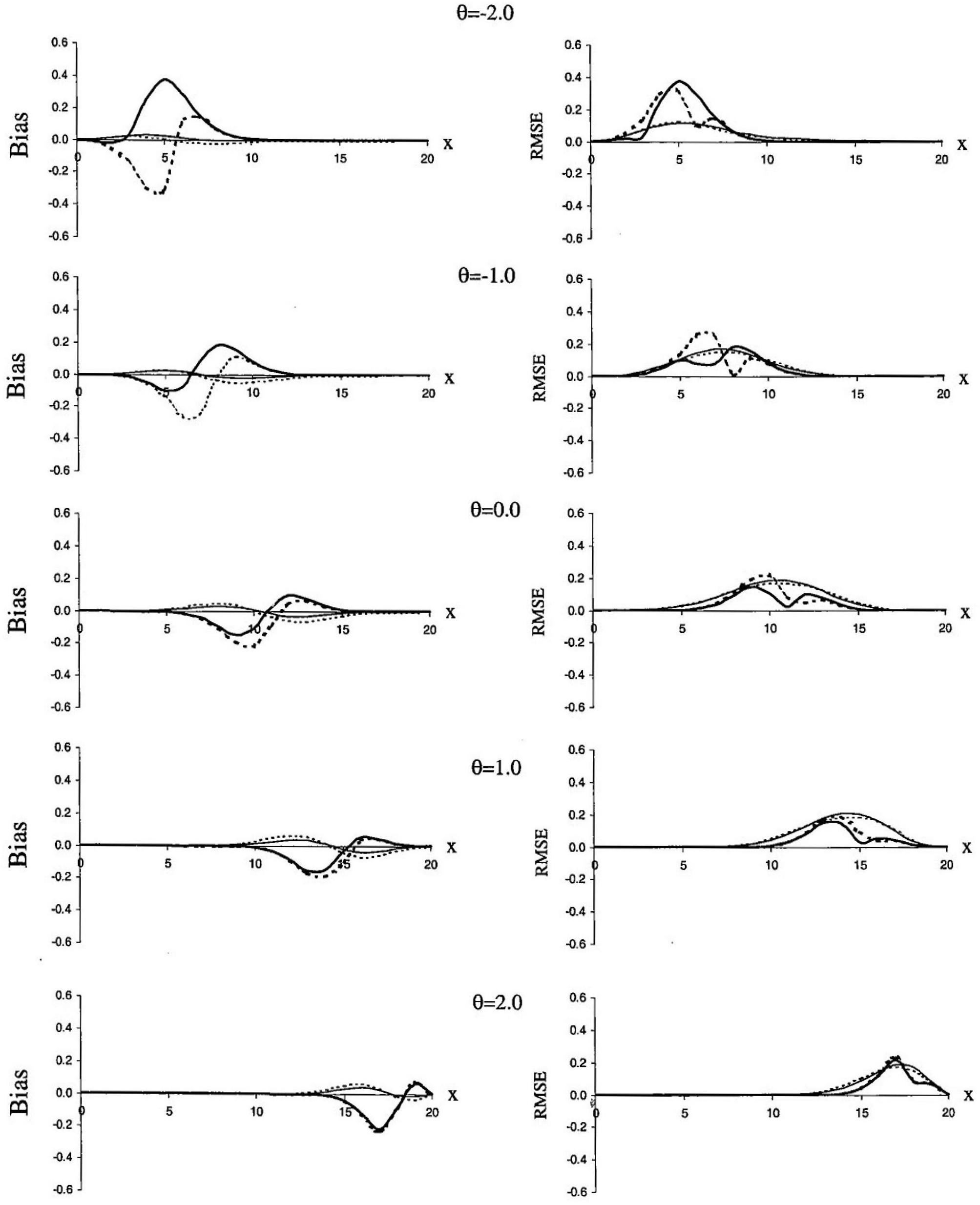

FIGURE 4. *Bias and RMSE functions of marginal equating method (bold solid line),TCF method (bold dotted line), estimated conditional method (thin solid line), and posterior expected conditional method (thin dotted line) for θ = –2.0, –1.5, ..., 2.0 and n = 20*

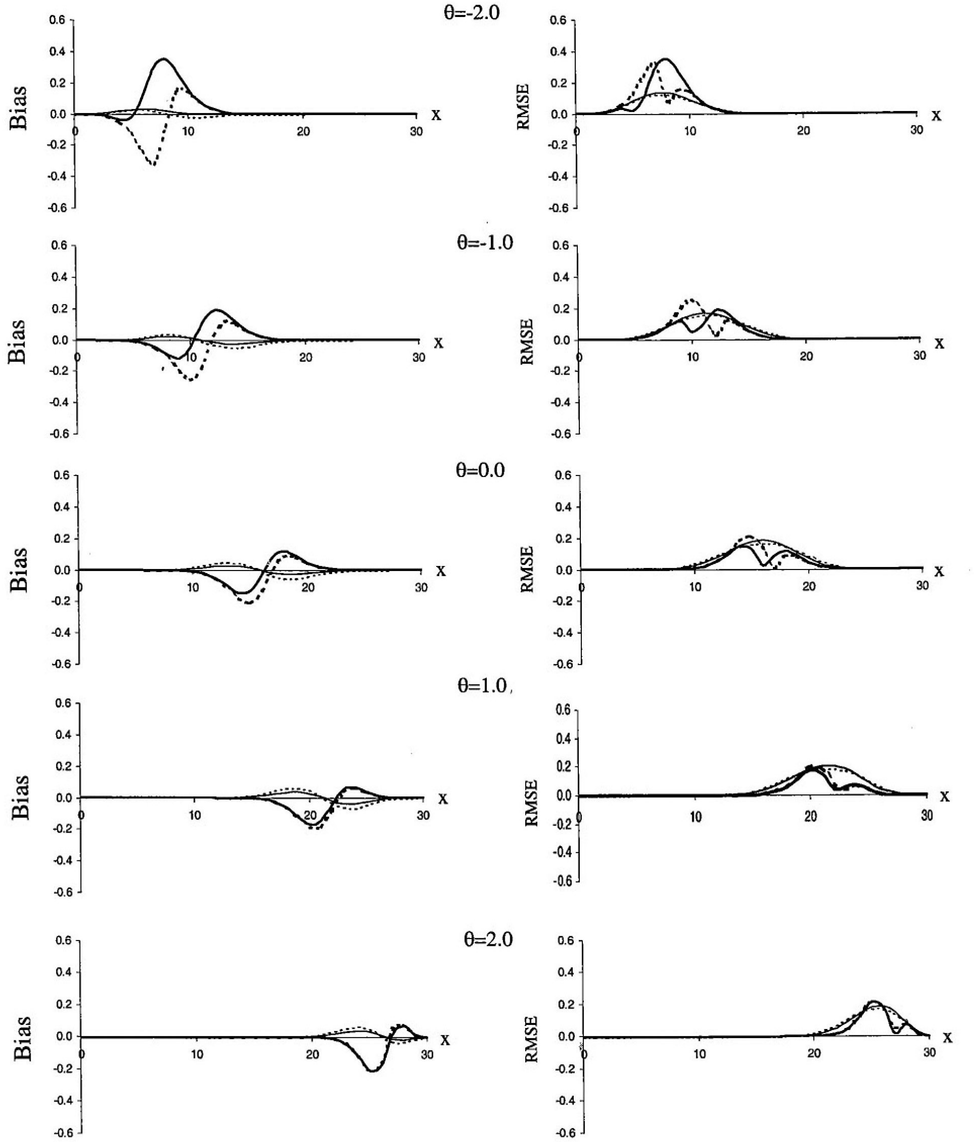

FIGURE 5. *Bias and RMSE functions of marginal equating method (bold solid line),TCF method (bold dotted line), estimated conditional method (thin solid line) and posterior expected conditional method (thin dotted line) for θ = –2.0, –1.5, ..., 2.0 and n = 30*

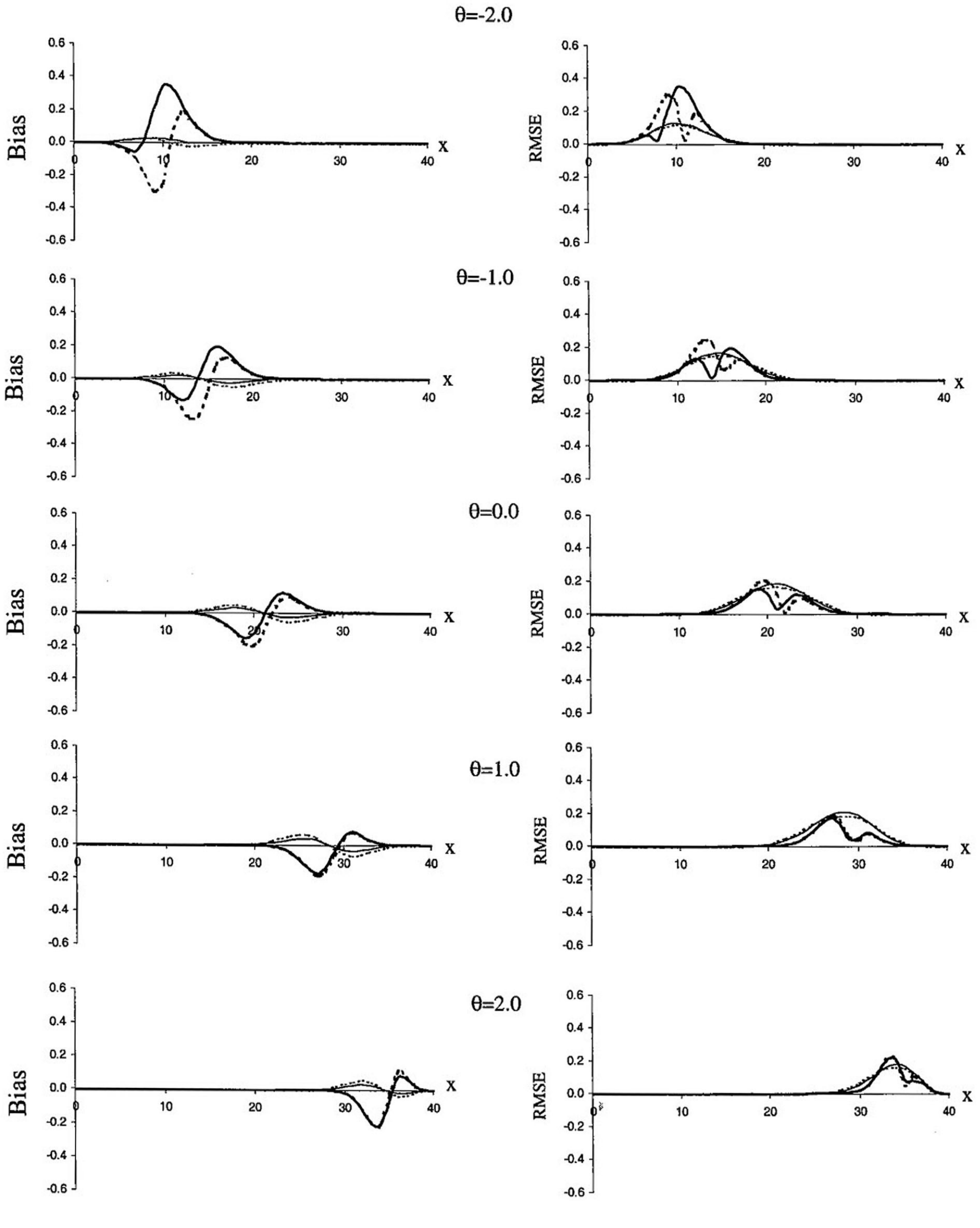

FIGURE 6. *Bias and RMSE functions of marginal equating method (bold solid line),TCF method (bold dotted line), estimated conditional method (thin solid line) and posterior expected conditional method (thin dotted line) for θ = –2.0, –1.5,..., 2.0 and n = 40*

The RMSE functions for the marginal and TCF method are obtained by taking the absolute value of their bias functions. The RMSE functions for the two conditional methods reflect both their bias and their variation over CAT administrations. For the negative *θ* values, the RMSE functions for the conditional methods are generally lower than for the marginal and TCF method. For the positive *θ* values, they are of comparable order. The reason for this increase in mean-squared error is that the relatively less accurate CAT scores at the higher *θ* values not only introduce the tendency to a flatter number-correct score distribution on the linear test, but also to more variation between these distributions over replicated CAT administrations.

Final Remarks

The results from this evaluation study can be summarized by observing that, for the lower θ values, the two conditional equating methods outperform the marginal and the TCF method both in bias and accuracy, whereas for the higher values, they outperform these methods in bias but showed comparable inaccuracy.

As explained earlier, the reason for the often large bias in the marginal and TCF method is their use of a single equating transformation for all examinees, which is a compromise of the transformations needed for the different observed-score distributions on the linear test for different ability levels. The two conditional methods take such differences into account and are virtually free of bias. Because the marginal and TCF transformations are fixed, CAT replications do not add any additional error. But the conditional transformations are random over these replications and, as a consequence, show less accuracy. For the lower θ values, the net result of both types of error is much in favor of the conditional equating methods, but for the higher values the choice depends on a preference for bias or inaccuracy. The trade-off between bias and inaccuracy is a fundamental law in statistics, and the results in this empirical study for the higher θ values just show another manifestation of it.

Of course, these conclusions cannot be generalized to any item pool, CAT algorithm, or reference test because they also depend on their features. Also, from a practical point of view, the comparison is not yet conclusive because we treated both the marginal equating transformation and the item parameter values as known, whereas in practice they have to be estimated.

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.

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

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

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.). *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). Washington, DC: American Psychological Association.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer et al., *Computerized adaptive testing; A primer* (chapter 5). Hillsdale, NJ: Erlbaum

van der Linden, W. J. (2000a). Constrained adaptive testing with shadow tests. In W J. van der Linden & C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 27-52). Boston: Kluwer.

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

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

van der Linden, W. J. (submitted). *Evaluating equating error in observed-score equating*.

van der Linden, W. J. (in preparation). *A Bayesian approach to observed-score equating*.

van der Linden, W. J., & Luecht, R. M. (1998). Observed-equating as a test assembly problem. *Psychometrika, 63*, 401–418.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

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