Local Observed-Score Equating With Anchor-Test Designs

Wim J. van der Linden¹ and Marie Wiberg²

Abstract

For traditional methods of observed-score equating with anchor-test designs, such as chain and poststratification equating, it is difficult to satisfy the criteria of equity and population invariance. Their equatings are therefore likely to be biased. The bias in these methods was evaluated against a simple local equating method in which the anchor-test score was used as a proxy of the proficiency measured by the test and the equating was conditional on this score. The results showed substantial bias for the two traditional methods under a variety of conditions but much smaller bias for the local method. In addition, unlike the traditional methods, the local method appeared to be quite robust with respect to changes in the difficulty and accuracy of the two tests that were equated. But like these methods, it appeared to be sensitive to a decrease in the accuracy of the anchor test as a proxy of the ability measured by the tests.

Keywords

anchor-test design, bias in equating, chain equating, equipercentile equating, local equating, poststratification equating

An important distinction between designs for data collection in test equating is between equivalent-group and anchor-test designs. The former allow us to derive the equating from the fact that the two tests are administered to a single group of test takers or two distinct groups sampled from a single population. For the latter, the equating is established through an anchor test that is administered to both groups. Reviews of the specific forms in which these two types of designs emerge in the practice of test equating are given in Angoff (1982), Kolen and Brennan (2004, section 1.4), and von Davier, Holland, and Thayer (2004, Chapter 2).

Some of the equating methods in the literature focus on full observed-score distributions, others only on their first few moments. The former are known as equipercentile methods in the testing industry. These methods equate the quantiles of observed-score distributions on the two tests. Examples of the latter are the linear equating methods, which equate the mean and variance of the distributions assuming that all higher order moments are already close enough to ignore their differences.

¹CTB/McGraw-Hill, Monterey, California
²Umeå University, Sweden

Corresponding Author:
W. J. van der Linden, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940
Email: wim_vanderlinden@ctb.com
The transformation used in equipercentile equating is a general statistical tool for mapping one arbitrary continuous distribution onto another (e.g., Wilk & Gnanadesikan, 1968). Suppose $U$ and $V$ are two arbitrary continuous random variables with realizations $u$ and $v$. Their distributions are described by the (cumulative) distribution functions $F_U(u)$ and $F_V(v)$. If $F_U(u)$ and $F_V(v)$ are strictly monotonically increasing, the transformation $u = \varphi(v)$ that maps the distribution of $V$ onto that of $U$ is

$$
\varphi(v) = F_U^{-1}(F_V(v)).
$$

(1)

The fact that the transformation equates the quantiles in the two distributions becomes more obvious when Equation 1 is presented as a set of parametric equations

$$
\alpha = F_U(u),
$$

and

$$
\alpha = F_V(v),
$$

(2)

where $\alpha \in [0, 1]$ is the cumulative probability shared by the quantiles $u$ and $v$ in the distributions of $U$ and $V$ (Lord, 1980, section 13.1).

The research in this article is on equipercentile equating for data-collection designs with an anchor test. Specifically, two existing methods of anchor-test equating are compared with a new local equating method for the same type of design. As explained below, methods of local equating have been introduced to deal with the fact that the existing observed-score equating methods have difficulty satisfying the requirements of equity and population independence (van der Linden, 2000, 2006a, 2006b, 2010). Equity means that for any ability level, the distributions of the equated scores and the scores to which they are equated are indistinguishable (Lord, 1980, section 13.5). Although equity is generally acknowledged to be a fundamental requirement, the equating literature has shown little progress in dealing with it. The criterion of population independence requires equating transformations to remain the same with any change in the populations of test takers. This requirement has led to several attempts at assessing the degree of population dependence of a given equating as well as studies of differences in population dependence between equating methods (for a recent review of the results, see the special issue of this journal edited by von Davier & Liu, 2008) but not to any proposal as to how existing population dependence can be reduced.

**Equipercentile Equating With Anchor-Test Designs**

The two main methods of equipercentile equating for an anchor-test design are the chain-equating and poststratification methods (Kolen & Brennan, 2004, Chapter 5; von Davier et al., 2004, Chapters 10 and 11). Both methods use estimates of population distributions of the observed scores on the two tests but require different assumptions to arrive at the estimates. The key features of the methods and their assumptions are briefly reviewed.

Let $X$ be the original version of a test and $Y$ a new version built to the same specifications. For notational convenience, the two tests are assumed to be of equal length, $n$. In addition, $Y$ and $X$ are used to denote the number-correct score on the two tests. It is assumed that $Y$ has to be equated to $X$. The transformation that performs the equating maps the possible values $Y = y$ to $X = x$. This mapping is denoted as $x = \varphi(y)$. (If scaled scores instead of number-correct scores are reported for $X$ and $Y$, the transformation has to be followed by a standard monotone transformation, but the basic equating is still between the number-correct scores.)
To establish the transformation, the scores of the two groups of test takers on an anchor test, A, are used. A will be used to denote the score on the anchor test, a for a possible value of this score, and \( A \) for the entire set of possible values. Throughout this article, for convenience, it is assumed that score A is neither part of \( X \) nor \( Y \). This assumption does not lead to any loss of generality; if \( A \) is a common part of \( X \) and \( Y \), the unique parts of these scores can just be equated using the methods in this article and the score on A added to the result. Because the two groups are not equivalent, it is assumed that they are sampled from different populations \( P \) and \( Q \), with distribution functions denoted as \( F_P^X(x) \) and \( F_Q^Y(y) \), respectively. Likewise, the distribution functions for the anchor test in the two populations are denoted as \( F_P^A(a) \) and \( F_Q^A(a) \). For ease of exposition, each of these functions is taken to be continuous and strictly monotone.

**Chain Equating**

The transformation in Equation 1 can be used both to equate \( Y \) to \( A \) on population \( Q \) and \( A \) to \( X \) on population \( P \). For the equating of \( Y \) to \( X \), it seems natural to use a composition (or chain) of these two transformations. However, this procedure would be invalid because the distribution functions in the two separate transformations are not defined over a common population.

The problem is generally acknowledged by making the following two assumptions:

1. The equating from \( Y \) to \( A \) is population invariant. That is,
   \[ a = F_Q^{-1}(F_{QY}(y)) = F_A^{-1}(F_Y(y)). \]  
   \( \text{(3)} \)

2. Likewise, the equating from \( A \) to \( X \) is population invariant,
   \[ x = F_P^{-1}(F_PA(a)) = F_X^{-1}(F_A(a)). \]  
   \( \text{(4)} \)

When these two assumptions hold, the transformation for chain equating becomes

\[ x = \varphi^{\text{chain}}(y) = F_P^{-1}(F_PA(F_Q^{-1}(F_{QY}(y))))). \]  
\( \text{(5)} \)

The transformation can be estimated by substituting sample distribution functions for the population functions in Equation 5.

**Poststratification Equating**

This type of equating is also hampered by the differences between the score distributions for \( P \) and \( Q \). To resolve this issue, Braun and Holland (1982, section 3.3.2) introduced the notion of a synthetic population

\[ T = wP + (1 - w)Q, \]  
\( \text{(6)} \)

with \( w \) an arbitrary weight, and suggested estimating the distributions of \( X \) and \( Y \) for this special population. The equating from \( Y \) to \( X \) is then the transformation in Equation 1 for the distribution functions for the two tests defined on \( T \). These distribution functions can be derived if the following two assumptions are satisfied:

1. The conditional distributions of \( X \) given \( A = a \) are population invariant. That is,
   \[ F_{PX|A}(x) = F_{X|A}(x), \text{ for all } a \in A. \]  
   \( \text{(7)} \)
2. Likewise, the conditional distributions of $Y$ given $A = a$ are population invariant,

$$F_{Q|a}(y) = F_{Y|a}(y), \text{ for all } a \in A.$$  \hspace{1cm} (8)

Using a total probability formula, the distribution functions of $X$ and $Y$ on $T$ can be written as

$$F_{TX}(x) = \sum_{a=1}^{m} [wP(a) + (1-w)Q(a)]F_{X|a}(x)$$  \hspace{1cm} (9)

and

$$F_{TY}(y) = \sum_{a=1}^{m} [wP(a) + (1-w)Q(a)]F_{Y|a}(y),$$  \hspace{1cm} (10)

where $P(a)$ and $Q(a)$ are the probability functions for the scores on anchor-test $A$ in $P$ and $Q$.

The equating transformation for poststratification is

$$x = \phi^{\text{post}}(y) = F_{TX}^{-1}(F_{TY}(y)).$$  \hspace{1cm} (11)

It can be estimated by choosing weight $w$ and substituting sample distribution functions for their population equivalents in Equations 9 and 10.

To get stable estimates of the population distribution functions in the chain and poststratification equatings, it is common practice to smooth the sample distributions before estimating the equating transformations. A very flexible family of models for smoothing observed-score distributions is the family of loglinear models proposed in Holland and Thayer (2000). The practical necessity of this presmoothing (or, alternatively, postsmoothing) is acknowledged, but because the main focus of this article is on a comparison between the nature, assumptions, and behavior of different equating methods for anchor-test designs, a discussion of its technical aspects is omitted.

Criteria of Equating

Lord (1980, section 13.5) formulated three fundamental criteria against which all equating methods should be evaluated: (a) equity, (b) population invariance, and (c) symmetry. The first criterion requires the equated score $\phi(Y)$ and the score on the original test, $X$, to be identically distributed for each test taker. This criterion summarizes the goal of observed-score equating; it is required that anyone with an interest in the test scores be indifferent as to whether $X$ or $Y$ is taken. The second criterion requires the transformation to be the same regardless of the population used to derive it. The last criterion, which was introduced by Lord mainly to emphasize the differences between the equating of $Y$ to $X$ and the regression of one score on the other, requires the transformation to be independent of which test form is labeled as $X$ or $Y$.

In addition, it is easy to overlook the fact that equating only makes sense if the two tests measure the same variable. This criterion of an identical measurement construct can be taken as the requirement that the distributions of the observed scores of the individual test takers on tests $X$ and $Y$ should be ordered identically (van der Linden, 2000). Finally, von Davier et al. (2004, section 1.1) added the criterion of equal reliability to the list. For a discussion of these criteria, as well as some weaker versions thereof, see, for instance, Harris and Crouse (1993), Kolen and Brennan (2004, section 1.3), and Yen (1983).

It is clear that the difficulty of satisfying the criteria of equity and population invariance is the Achilles’ heel of chain and poststratification equating. Neither of the methods is based on an explicit attempt to match the equated scores $\phi(Y)$ and the original scores $X$ for the individual
test takers. Besides, through Equations 3 to 4 and Equations 7 to 8, both methods simply assume population invariance but do not have any built-in check of this assumption, let alone a safeguard against violations of it.

Also, it is difficult to buy the notion of a synthetic population in Equation 6 as the target population for the equating. The notion seems motivated by the idea of two-stage sampling from a population with strata \( \mathcal{P} \) and \( \mathcal{Q} \) and weights \( w \) and \( 1 - w \). However, weighted sampling does not occur in the typical equating study with an anchor-test design. In fact, in real-world applications, the equating transformation is used only for the scores of the examinees that take the new test, \( Y \). From the population-based perspective that underlies the poststratification method, choosing a target population that also involves the old population seems therefore inappropriate.

The method of chain equating may be sensitive to propagation of estimation errors. Its transformation involves a chain of four estimated distribution functions. Although not documented in the research literature, it seems reasonable to assume that random estimation errors in the earlier functions, or systematic errors due to the use of a smoothing method, will be multiplied by later errors. If this happens, the results may be dangerous, especially for errors in the tails of the distributions, which are always more difficult to estimate or smooth.

The alternative method of observed-score equating in the next section avoids the assumption of the test takers’ being sampled from (possibly different) populations entirely. Instead, it focuses on the scores of the individual test takers and tries to optimize the equity of equated scores \( \phi(Y) \) by matching them directly with their corresponding scores on \( X \).

**Local Equating**

Suppose the number-correct score \( Y_p \) by a single test taker \( p \) on a new test form is to be equated to his or her (unknown) score \( X_p \) on an old form. A basic assumption in any test theory is that both scores have random error and therefore vary across replications. Thus, \( X_p \) and \( Y_p \) should be conceived of as two different random variables, each with its own distribution. Let \( F_{X_p}(x) \) and \( F_{Y_p}(y) \) be the distribution functions of the two scores. From Equation 1, it follows immediately that the transformation that maps the distribution of \( Y_p \) onto \( X_p \) is

\[
x = \varphi_p(y) = F_{X_p}^{-1}(F_{Y_p}(y)).
\]  

(12)

This transformation always satisfies the criterion of equity for the test taker. It also satisfies the assumption of population invariance—no matter what population the test taker would be sampled from, the use of Equation 12 would always lead to the same equating for his realized score \( Y_p = y \).

Continuing the thought experiment, suppose that we now have to deal with a second test taker, \( q \), as well. We are then faced with a basic choice between (a) using the individual transformations for \( p \) and \( q \) separately or (b) considering these two test takers as a population, calculating its distribution functions for \( X \) and \( Y \), and performing the equipercentile transformation on these functions.

Obviously, the first option still guarantees equity for the equatings for \( p \) and \( q \). Also, it does not involve any population dependency. But both features would be lost if we choose the second option. Assuming simple random sampling of \( p \) and \( q \), the distribution function for the population \( \{p, q\} \) on test \( X \) is the average of \( F_{X_p}(x) \) and \( F_{X_q}(x) \), say \( F_X^*(x) \). The distribution function for the same population on test \( Y \) is similarly defined as \( F_Y^*(y) \). Therefore, the equating transformation for the second option becomes

\[
x = \varphi(y) = F_X^{-1}(F_Y^*(y)).
\]  

(13)
The transformation is a compromise between the two transformations that would actually be needed to guarantee equity of the equatings. Also, any change of the population, for instance, the addition of a third person, would immediately result in another transformation as a new compromise.

However, although the choice between the two options has a clear winner, this was a thought experiment and we have no direct access to the score distributions of the individual test takers that are required to estimate the equating transformations in Equation 12. Local equating (van der Linden, 2000, 2006a, 2010) is based on the idea that it is nevertheless better to approximate these individual distributions using whatever information is available about the test takers than to compromise between them. The set of transformations for the individual test takers calculated from these approximations is expected to be closer to their true equating transformations than any single transformation based on the two populations that took X and Y, even though the estimation of the distribution functions for these two populations may be statistically more straightforward. The gain is expected to be greater (a) the greater the variation between the proficiencies of the test takers and (b) the better we succeed in approximating their individual observed-score distributions.

Test theory rests on the general assumption that when a test is fixed, the differences between the distributions of the observed scores of individual test takers are determined entirely by their proficiencies. This assumption underlies, for instance, the binomial-error model as well as the attempts in classical test theory to define conditional standard errors of measurement. It also underlies item response theory (IRT); for its models, it holds that when the item parameters are fixed, the differences between observed-score distributions depend entirely on the person parameters.

One way of approximating the transformations in Equation 12 is therefore to find a proxy for the (common) proficiency measured by the two tests and use the conditional distributions of X and Y given the proxy to establish the equating transformations for the individual test takers. Let \( t \) denote such a proxy. (The term “proxy” is used instead of “estimate” because, due to the conditioning in Equation 12, the scale on which the proficiency is represented does not matter.) The equating transformations for the individual test takers would then take the form of the family of transformations

\[
x = \varphi_t(y) = F_{X|t}^{-1}F_{Y|t}(y), \quad t \in \mathcal{R},
\]

where \( F_{X|t}(x) \) and \( F_{Y|t}(y) \) denote the distribution functions of the conditional distributions of \( X \) and \( Y \) given \( t \), respectively, and \( \mathcal{R} \) denotes the range of possible values of \( t \).

Of course, the challenge for local equating is to identify reasonable proxies for the proficiency measured by the tests and find satisfactory estimates of the conditional distributions of \( X \) and \( Y \) given the proxy. One example is the equating of an adaptive test \( Y \) to a linear reference test \( X \) by van der Linden (2006b). Because adaptive tests are run from an item pool calibrated under an IRT model, the family of equating transformations can be based on the conditional distributions of \( X \) and \( Y \) given \( \theta \). These transformations can simply be calculated using the well-known recursive algorithm by Lord and Wingersky (1984) as soon as the items have been calibrated. During operational testing, we only have to estimate \( \theta \) from the test taker’s response vector on the adaptive test \( Y \) and use the transformation at this estimate to equate the score on \( Y \) to a score on the reference test.

For the anchor-test design studied in this article, an obvious choice of proxy for the proficiency measured by the two tests is the observed score on the anchor test. Then the conditional distributions of \( X \) and \( Y \) given \( A = a \) are used to define the family of transformations

\[
x = \varphi_a(y) = F_{X|a}^{-1}(F_{Y|a}(y)), \quad a \in \mathcal{A},
\]
and use this family as an approximation to the (unknown) true family of equating transformations. In an actual equating, in order to use the family, the score $A = a$ on the anchor test should be known for the test taker to pick the member $x = \varphi_a(y)$ from Equation 15 for the equating of his or her observed score $Y = y$ to the score on test $X$. This routine thus requires the anchor test to be administered along with $Y$ or as an internal section of it.

To allow for a comparison with the assumptions underlying traditional chain and poststratification equating, the (only) assumption involved in this choice of family is formulated explicitly as:

1. The observed score on the anchor test, $A$, is a good proxy of the common proficiency measured by tests $X$ and $Y$.

The assumption implies two different criteria for the anchor test: (a) measurement of the same proficiency as $X$ and $Y$ and (b) accuracy. Generally, test forms measure the same proficiency if the observed-score distributions of the test takers on these forms display identical stochastic order. Joint fit of a mainstream response model as the three-parameter logistic (3PL) model is sufficient for this criterion to hold (van der Linden, 2000, propositions 1 and 2). The current study does not assume any specific response model to hold, so it has to resort, for instance, to a check on the ordinal correlations between $A$ and $X$ and $Y$. The results of such checks will be more favorable the better the composition of the anchor test mirrors that of the tests that are equated—a well-known recommendation for equating in an anchor-test design (e.g., Cook & Petersen, 1987). The second criterion of accuracy is required because the true proficiencies of the test takers are to be approximated as closely as possible. The criterion implies a straightforward check on the reliability of the anchor test—a feature primarily controlled by its length and the discriminating power of its items. The impact of both parameters on equating error is addressed in the empirical study reported below.

**Comparison Between Local and Traditional Equating**

Each of the three alternative equating methods uses different marginal and/or conditional distributions to be estimated in a study with an anchor-test design:

1. In the chain-equating method, the equipercentile transformation is derived from the marginal population distributions of $X$ and $Y$ as well as the marginal population distributions of the scores on $A$ for the two populations that took $X$ and $Y$.

2. In the poststratification method, the conditional distributions of $X$ and $Y$ given $A = a$ are used to derive the marginal distributions of $X$ and $Y$ for a synthetic population, $T$. The equipercentile transformation is then derived from these two distributions.

3. In the local equating method, the conditional distributions of $X$ and $Y$ given $A = a$ are used directly to derive a family of equipercentile transformations for the individual test takers.

A reader with a more population-based perspective might indicate that this local method of equating is in fact based on the same two assumptions of population invariance as the poststratification method in Equations 7 and 8 because both use the same conditional distributions. This study has a different view but acknowledges the subtlety of the issue. If $A$ is the actual proficiency measured by the tests, both the distributions of $X$ and $Y$ depend on $A = a$ only. Hence, conditioning the scores $X$ and $Y$ on $A = a$ makes them automatically population independent. Just as the distributions of $X$ and $Y$ given $\pi$ are population independent in the case of a binomial-error model, or, for that matter, the distributions given $\theta$ for an IRT model (local independence). If $A$ is a less-than-perfect proxy, test takers with neighboring proficiencies have a substantial probability of receiving the same observed score on $A$. As a result, conditioning on $A = a$ implies a mixing of the observed-score distributions of test takers. However, as just noticed,
anchor tests are better the closer they are selected to the content specifications of the tests to be equated. It is therefore expected that the differences between the proficiencies of test takers with the same score $A = a$ are small, especially relative to those in the populations $P$ and $Q$ in the poststratification method, for which the conditional distributions of $X$ and $Y$ given $A = a$ are mixed over all possible values $a \in A$; see Equations 9 and 10.

The idea that test equating should be based on score distributions adjusted for the proficiencies of the populations that take the test forms is omnipresent in the equating literature. For example, von Davier et al. (2004, p. 2) observe that equating always requires some kind of control of the differences in abilities between the persons that take the two tests. Other examples are found in an entire special issue of *Applied Measurement in Education* (Dorans, 1990) devoted to the topic of matching the samples that take the old and new test forms to give them identical proficiency distributions before the equating transformation is inferred. Interestingly, the majority of the contributions to this special issue use the scores on the anchor tests to perform the matching. The idea of local equating takes this logic of matching one step further: Rather than attempting to get identically distributed proficiencies for the groups that take the old and new test forms, local equating attempts to equate the observed-score distributions of individual test takers with identical proficiencies on the two forms.

Of course, a conspicuous difference between traditional and local equating is the use of a single transformation versus a family of transformations with the proficiency levels of the test takers as the index. At first sight, it may seem to be embarrassing to have to assign different equated scores to test takers with the same score $Y = y$. But actually the opposite practice of automatically assigning the same equated score to such cases easily leads to fairness problems. An example in van der Linden (2010) illustrates this point (the same reference should be consulted for several other motivations of local equating): Test takers seldom reach the same number-correct score on $Y$ by answering the same items correctly. It is therefore easy to run into cases in which some of them realize the same score by producing correct answers to more difficult items than the others. Technically, this means realized observed scores for the former from the lower tail of their observed score distributions but from the upper tail for the latter. Bias in traditional equating is a direct consequence of ignoring this information and using the same transformation for all test takers. Conditioning on all the information about the test takers’ proficiencies available in the equating study helps to adjust for this bias.

It may take some time for the necessity of this adjustment to become generally accepted. On the other hand, the history of testing shows an identical development for the use of the standard error of measurement. The traditional standard error was a single quantity reported for all test takers in a population. It is now known that this error is a biased estimate of the true individual standard errors for the test takers, and it has become customary to report numbers as closely as possible to these true individual errors. The true quantities approximated are the standard deviations of the same conditional observed-score distributions as those in the definition of the family of equating transformations in Equation 14.

**Empirical Evaluations**

The goal of the empirical part of this research was to evaluate the chain, poststratification, and local equating methods. Specifically, this study aimed to assess the bias in these three methods against equating based on transformations that are known to be true. Although the equating literature has had an interest in the standard error of equating for a long time, the present authors believe that absence of any bias is a more fundamental criterion of success in equating. The mere goal of equating is to remove the bias in the scores on a new test form as a substitute of the test takers’ scores on an old form, and any bias left in the new scores—or extra bias added to them in
the process of the equating—would invalidate the equating. For the sake of completeness, however, the root mean square error (RMSE) of the three methods was also evaluated.

The equating was between two different 40-item test forms randomly sampled from two different sections of a previous item pool from the Law School Admission Test (LSAT) with a third form as the anchor test. In addition to an evaluation of the bias and RMSE of the three methods for this equating, their robustness was evaluated with respect to variations in the composition of test forms and the anchor tests. More specifically, the robustness was evaluated varying the following factors:

1. difficulty of Y relative to X,
2. accuracy of Y relative to X,
3. difficulty of anchor test A relative to Y and X,
4. accuracy of A relative to Y and X, and
5. length of A.

Method

Because the family of true equation transformations is generally unknown for real-world tests, the three methods were evaluated using simulated data for the tests in this study. More specifically, data were simulated under the 3PL response model. For this case, the observed-score distributions of the simulated test takers given their proficiencies are generalized binomials and can easily be calculated from the item parameters using the earlier mentioned algorithm by Lord and Wingersky (1984). From these distributions, the family of true equating transformations derived from Lord’s requirement of equity can then be calculated as

$$x = \psi_\theta(y) = F_{X|\theta}^{-1}F_{Y|\theta}(y), \quad \theta \in \mathbb{R}.$$  \hfill (16)

Observe that the 3PL response model is adopted here for evaluation purposes only. It is not necessary to assume the model to hold in any actual application of any of the equating methods in this article.

Bias for the three equating methods is defined as the difference between their actual transformations and the true transformations \(\psi_\theta(y)\) in Equation 16. Observe that there is a different bias for each possible score \(Y = y\); in other words, bias should be treated as a function of \(y\). Inasmuch as there is a different true transformation for each true proficiency \(\theta\) in this study, a different bias function was obtained for each \(\theta\). The same holds for the RMSE functions. The local equating transformations also depend on the actual score on the anchor test, \(A = a\). Because \(A\) is random, the bias function for these transformations is evaluated taking the expectation over \(A\) given \(\theta\):

Thus, formally the bias functions are defined as

$$\text{bias}(\psi^{\text{post}}(y); \theta) = \psi^{\text{post}}(y) - \psi_\theta(y), \quad \theta \in \mathbb{R},$$  \hfill (17)

$$\text{bias}(\psi^{\text{chain}}(y); \theta) = \psi^{\text{chain}}(y) - \psi_\theta(y), \quad \theta \in \mathbb{R},$$  \hfill (18)

$$\text{bias}(\psi_a(y); \theta) = \mathcal{E}_{\theta|\phi} [\psi_a(y) - \psi_\theta(y)], \quad \theta \in \mathbb{R}.$$  \hfill (19)

Generally, the expectation of the expression in Equation 19 over the conditional distribution of \(A\) given \(\theta\) is not available in closed form but can be calculated using a Monte Carlo simulation. The response model is then used to randomly generate response vectors on X, Y, and the anchor test at \(\theta\), whereupon the expression in Equation 19 is averaged over the distribution of the number-correct scores on A given \(\theta\). This approach was followed in the current study.
Data

The items were previously operational items calibrated under the 3PL response model. The means and standard deviations of the two forms and the anchor tests are given in Table 1. The test characteristic curves (TCCs) of the three forms are given in Figure 1. These data show that form X was less difficult than form Y and slightly more discriminating. In addition, anchor form A was closer to X than to Y.

The response vectors on X, Y, and A were generated for $N = 50,000$ test takers randomly sampled from a normal distribution. For population $P$ and $Q$, $N(-0.5, 1)$ and $N(0.5, 1)$ were used, respectively. The difference between the population means is relatively large but the primary goal for this study was to demonstrate the range of possible effects, not to represent typical cases. The sample size was chosen to be large because this study’s interest was exclusively in accurate estimates of the bias functions for the equating methods. In particular, these estimates were not to be confounded with sampling error.

From the responses, the number-correct scores on the three test forms and the (conditional) distribution functions required for the chain, poststratification, and local equating methods were calculated. The distribution functions for the synthetic population in the poststratification method were calculated for $w = .5$, which is a common choice for equally sized or equally relevant populations (Kolen & Brennan, 2004, section 4.1.5). To calculate the expected value in the bias function in Equation 19, the response vectors were replicated 10 times for each of the 50,000 test takers. As the scores were discrete, all equipercentile transformations were found using simple linear interpolation (Kolen & Brennan, 2004, section 2.5). As the true equating transformation, the closest transformation at $y = -2.0, -1.9, \ldots, 2.0$ in Equation 16 was chosen for the randomly sampled $\theta$ for each test taker. Observed scores with relative frequencies smaller than $10^{-4}$ for some values of $y$ were omitted from the study. In practice, only a few test takers would have scores at the omitted values and their equated scores would have been too inaccurate to be reported. As a result, some of the error functions in the figures in the next section do not run over the entire range of possible values of $y$ for all values of $\theta$.

The conditions with the variation in the difficulty and accuracy of the test forms were created as follows: The relative difficulty of the two standard forms was manipulated by adding and subtracting 0.5 from the difficulty parameters of the items in Y. The relative measurement accuracy of the two forms was manipulated by multiplying the discrimination parameters of Y by 0.5 and 2.0. The same same procedure was used to realize conditions with a more and less difficult as well as a more or less accurate anchor test. Again, these variations were chosen to demonstrate the range of the effects of these parameters on the bias in the equating; typical cases are within these ranges.

In addition, the length of the anchor test was shortened by 25% and 50%. To prevent confounding between the length and the composition of the anchor tests, the shorter tests were
randomly sampled from the standard anchor test. Generally, the length of the anchor tests fits those in earlier reviews of empirical studies with an external anchor design, which ranged from 20 to 60 items (e.g., Kolen & Brennan, 2004, p. 271; Petersen, Marco, & Stewart, 1982; von Davier et al., 2004, p. 156). Fitzpatrick (2008) warns explicitly against anchor tests with fewer than 15 items.

Finally, observe that tests X and Y were chosen to be smaller than typical operational tests. This conservative choice allows us to safely generalize the results to longer tests.

Results

To reduce the overload of information, only the results for $\theta = -2, -1.5, \ldots, 2$ are reported. All results for the omitted $\theta$ values did fit the general pattern discussed in this section.

EQUATING TRANSFORMATIONS (STANDARD CASE). Figure 2 shows the transformations for the equatings of the two standard forms. The true transformations are generally ordered by their value of $\theta$ (i.e., the larger the value of $\theta$, the higher the equated score $\varphi(y)$). Observe that there is one transformation for the chain and poststratification methods and that the transformation for the chain equating yielded systematically lower equated scores than that for the poststratification method. (The same difference was observed for all conditions in this study.) But for the true and local methods there is one transformation for each value of $\theta$ and $\alpha$, respectively. The differences between the local transformations tended to be a little smaller than between the true transformations. Also, the local transformations were less smooth than the transformations for the other two methods because of sampling variation. This difference
reflects the fact that the estimates of the conditional distributions of $X$ and $Y$ given $a$ are based on smaller subsets of the data than the estimates of the marginal distributions in the chain and poststratification methods, especially for the more extreme values of $a$. Use of a smoothing technique would immediately have given them a polished look. In this study, this option, which is standard practice in real-world equating, was omitted to avoid possible confounding of different sources of bias.

**Bias functions (standard case).** The three remaining plots in Figure 2 show the bias functions for the three equating methods. For the method of chain equating, the bias was extremely large and negative. For the method of poststratification, it was somewhat smaller but still substantial. These results are explained by the differences between the transformations for the two methods relative to the true transformations (see the two top panels of Figure 2). The bias for the local method was much smaller and slightly on the positive side. The tendency to positivity was not reproduced for the other conditions in this study, so it was not expected to be systematic.

**Robustness of equating.** The next four figures display the results from the study of the robustness of the three equating methods with respect to changes in test form $Y$. For the more difficult version of $Y$ in Figure 3, the bias functions were entirely comparable to the standard case in Figure 2. But for the less difficult version in Figure 4, the variation between the true transformations increased and so did the variation between the bias functions for each of the three methods. For the two traditional methods, the change in accuracy of test form $Y$ in Figures 5 and 6 had a dramatic impact on the size of the bias, particularly for the combination of chain equating and the less accurate version of $Y$ in Figure 6. (Observe the difference between the scales of the vertical axes in Figures 5 and 6 and the other figures.) For the local method, the bias functions were quite
**Figure 3.** True, chain, poststratification, and local equating transformations as well as their bias for a more difficult test form $Y$

**Figure 4.** True, chain, poststratification, and local equating transformations as well as their bias for a less difficult test form $Y$
Figure 5. True, chain, poststratification, and local equating transformations as well as their bias for a more accurate test form $\gamma$

Figure 6. True, chain, poststratification, and local equating transformations as well as their bias for a less accurate test form $\gamma$
robust with respect to variation in the accuracy of Y and remained entirely comparable with the standard case in Figure 2.

As already indicated, although robust with respect to changes in the test that is equated, the local method is expected to be less robust with respect to changes in the anchor test because such changes may have a direct impact on the quality of its scores as a proxy of the true proficiency measured by X and Y. This expectation is confirmed by the results in Figures 7 through 12. The two traditional methods appeared to be quite robust with respect to any of the changes in the difficulty (Figures 7 and 8) or accuracy of the anchor test (Figures 9 and 10); their bias functions were entirely comparable to those for the standard case in Figure 2. The same holds for the robustness of the local method with respect to the changes in the difficulty of the anchor test (Figures 7 and 8). Both for the increase and decrease in difficulty, the bias remained small and was entirely comparable to the results in Figure 2. However, variation in the accuracy of the scores on the anchor test did have a serious impact on the equatings. As shown in Figure 9, for the more accurate anchor test, the bias for the local method vanished nearly completely. But for the condition with the less accurate anchor test in Figure 10, it was much larger and became comparable to that for the poststratification method. The same effects were observed for the reduction of the length of the anchor test. For the 25% reduction (Figure 11), the local method still outperformed the two traditional methods. But when its length was halved (Figure 12), the local method showed bias close to that for the poststratification method.

It should be observed that a decrease of the length of the anchor test (Figures 11 and 12) is statistically equivalent to a decrease of the accuracy of its scores (Figure 10). In both cases, A becomes a less reliable proxy for the proficiency measured by X and Y. Under this condition, conditioning on \( A = a \) in the local method involves a mixing of the observed-score distributions.
Figure 8. True, chain, poststratification, and local equating transformations as well as their bias for a less difficult anchor test

Figure 9. True, chain, poststratification, and local equating transformations as well as their bias for a more accurate anchor test
Figure 10. True, chain, poststratification, and local equating transformations as well as their bias for a less accurate anchor test

Figure 11. True, chain, poststratification, and local equating transformations as well as their bias for a 25% shorter anchor test (n = 30)
of the individual test takers over a larger range of their true proficiencies than is desirable (see above for a discussion of this topic). Consequently, the method has less opportunity to be local, and its performance begins approximating that of the equating methods based on full marginal
observed-score distributions (see the last section for a recent empirical confirmation of this explanation).

**RMSE of equating.** For completeness, this study also looked at the RMSE functions of the three equating methods. For the chain and poststratification methods, these functions are just those in Figures 2 through 12 with the negative part of the curves reflected about the horizontal axis. The local method does have additional random error due to the sampling from $A$ given $\theta$ (see Equation 19). However, the error appeared to be entirely negligible relative to the systematic component of the bias functions in Figures 2 through 12. Therefore only one example is given (Figure 13), which shows the RMSE function for the standard conditions in Figure 2.

**Discussion**

One conclusion from this evaluation study is general pessimism about the two traditional methods of anchor-test equating. Their results appeared to be considerably biased. Also, the shape of their bias functions appeared to be dependent on the proficiency of the test taker, which makes it generally difficult to correct the equatings for bias. Of these two methods, the chain equating method performed consistently worse than the poststratification method, a fact already attributed to the propagation of statistical error: Because of the chaining of four individual transformations, each next transformation operates on the errors in the previous transformations and it seems reasonable to expect an end result that involves accumulation of them.

The reason why these two traditional methods are biased is their one-size-fits-all approach to equating. Rather than approximating the equating transformations dictated by the observed-score distributions for the individual test takers, it produces a single transformation to compromise between all of them. The result is bias in the equating of each test taker. The same size of bias has been found under a variety of conditions in an earlier study of traditional equipercentile equating with an equivalent-group design (van der Linden, 2006a). As the bias is due to a structural aspect of traditional equating, it can be expected to show up in equating for any type of equating design; no matter how the equating population is defined, whether it is taken to be a single existing population, a synthetic population (Braun & Holland, 1982), or a reconstruction of a population based on matching (Dorans, 1990), the problem is just inherent in the notion of a population of test takers itself.

Another conclusion is that the local method performs much better provided the observed anchor-test scores meet certain minimum requirements as to their accuracy. If it does, it is practically bias free and the method clearly outperforms the two traditional methods in this respect. Also, the method turns out to be robust with respect to changes in the difficulty or accuracy of the test forms that are equated.

If the accuracy of the anchor test becomes too small, as when the size of the discrimination parameters in the anchor tests was reduced to half of their empirical size (Figure 10) or the length of the anchor test was halved (Figure 13), the anchor-test score loses its capability to identify the observed-score distributions of the individual test takers that are to be equated. The conditional distributions of $X$ and $Y$ given $A = a$ then no longer approximate these individual distributions satisfactorily and begin to resemble the marginal distributions on which the two traditional methods in this study are based. Empirical evidence of this explanation is offered in a recent study by Janssen, Magis, San Martin, and Del Pino (2009), who compared the local method in this article with a version in which the equating transformations were conditioned on ability estimates $\theta$ from the anchor test rather than its number-correct scores. Because of their higher accuracy, the $\theta$s produced less bias. However, unlike the current method, this alternative requires the three tests to jointly fit an IRT model.
Of the two factors that determine the accuracy of the anchor test (the size of the discrimination parameters of its items and its length), the former is an empirical issue beyond direct control of the testing agency. For professionally constructed tests, however, discrimination parameters half the size of those in the test forms used in this empirical study are expected to be unrealistic. The length of the anchor test, though, is selected by the testing agency. For the larger reduction in this study, the length of the anchor test \((n = 20)\) is exactly the lowest point of the range of 20 to 60 items found in the reviews of the empirical studies referred to earlier and just above the earlier warning against anchor tests shorter than 15 items by Fitzpatrick (2008). If traditional equating with anchor-test designs is to be improved, it also seems to be just the point above which such improvements become possible. A tradition of internal anchors in testing exists that seems to suggest using smaller numbers of items. But statistically only the number of items and their quality counts, not whether they are administered internally or externally.

It is interesting to note that the relative difficulty of the three test forms had no effect on the quality of the equating by the local method. This finding suggests that it should perform well in the context of vertical equating. The reason is no doubt its conditioning on the anchor-test score as a proxy of the test taker’s position on the ability scale.

It might be possible to improve the quality of the scores on the anchor test as a proxy of the proficiency measured by \(X\) and \(Y\). As already indicated, one approach would be to adopt weighted instead of simple number-correct scoring for the anchor test (although the use of weighting in other contexts has seldom led to convincing results). Another approach would be to use a regression technique to find the best common prediction of the anchor-test score from the two test forms that are equated and to condition on these predictions.

Another useful line of research is the application of smoothing techniques to local equating. The use of smoothing is common practice in all observed-score equating but has not yet been explored for local equating. Several techniques of presmoothing of the sample distributions are available, which could be applied directly to the distributions of \(X\) and \(Y\) given \(A = a\). A theoretically attractive form of presmoothing is offered by the kernel method of test equating (von Davier et al., 2004). If the sample distributions for the two tests differ mainly in their first two moments, linear equating becomes an attractive type of smoothing because it would only need data for the estimation of the conditional means and variances of \(X\) and \(Y\) given \(A = a\) (Wiberg & van der Linden, 2009). A complete review of the current smoothing options for observed-score equating is given in Kolen and Brennan (2004, Chapter 3).

Finally, observe that the 3PL response model was adopted here only for evaluation purposes in the empirical study. It is not necessary to assume the model to hold in any real-world application of the local equating method in this article. The only assumption made when using the method is that \(A\) is a satisfactory proxy of the true proficiency measured by \(X\) and \(Y\). Of course, the use of the response model in this evaluation study does not allow generalizing the findings to tests that do not fit it. But the restriction is expected to be not too dramatic because observed-score equating makes sense only if the tests measure the same unidimensional proficiency, and the 3PL model is one of the mainstream unidimensional models used to test this assumption in an operational context.

**Declaration of Conflicting Interests**

The authors declared no conflicts of interests with respect to the authorship and/or publication of this article.

**Funding**

The research in this article by Marie Wiberg was partially funded by the Swedish Research Council.
References


