Research Report

The Missing Data Assumptions of the Nonequivalent Groups With Anchor Test (NEAT) Design and Their Implications for Test Equating

Sandip Sinharay

Paul W. Holland

May 2008

ETS RR-09-16

The Missing Data Assumptions of the Nonequivalent Groups With Anchor Test (NEAT) Design and Their Implications for Test Equating

Sandip Sinharay and Paul W. Holland
ETS, Princeton, New Jersey

May 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The nonequivalent groups with anchor test (NEAT) design involves missing data that are missing by design. Three popular equating methods that can be used with a NEAT design are the poststratification equating method, the chain equipercentile equating method, and the item-response-theory observed-score-equating method. These three methods each make different assumptions about the missing data in the NEAT design. Though studies have compared the equating performance of the three methods under the NEAT design, none has examined the missing data assumptions and their implications for such comparisons. The missing data assumptions can affect equating studies because it is necessary to fill in the missing data or their distribution in some way in order to have a true, or criterion, equating function to compare the accuracy and bias of the different methods. If the missing data or their distribution are filled in using missing data assumptions that correspond to a given method, that may favor that method in any comparison with the others. This paper first describes the missing data assumptions of the three equating methods and then performs a fair comparison of the 3 methods using data from 3 different operational tests. For each data set, we examine how the 3 equating methods perform when the missing data satisfy the assumptions made by only 1 of these equating methods. The chain equating method is somewhat more satisfactory overall than the other methods in our fair comparison of the methods; hence, we recommend that equating practitioners seriously consider the chain equating method when using the NEAT design. In addition, we conclude that the results from the different equating methods will tend to agree with each other when proper equating conditions are in place. Moreover, to uncover problems that might not reveal themselves otherwise, it is important for operational testing programs to apply multiple equating methods and study the differences among their results.

Key words: Chain equating, poststratification equating, IRT observed-score equating, frequency estimation, raking, continuization

Acknowledgments

This research was supported by Institute of Education Sciences (IES) Unsolicited Grant R305U07009. Any opinions expressed in this paper are those of the authors and are not necessarily those of ETS or the IES. The authors thank Dan Eignor, Skip Livingston, and Tim Moses for their comments and Ayleen Stelhorn for the editorial help.

Table of Contents

List of Tables

Page

List of Figures

Page

1. Introduction

Among the various possible equating designs, the nonequivalent groups with anchor test (NEAT) design is widely used. The NEAT design is simultaneously one of the most flexible and one of the most complicated test-linking designs. In an application of the NEAT design, a new test (or test form) *X* is to be equated to an old test (or test form) *Y, X* is taken by a sample from population *P,* and *Y* is taken by a sample from a possibly very different population *Q*. In addition, an anchor test *A* is taken by both samples and allows one to study the difference in ability between *P* and *Q*. The pattern of data for the NEAT design is illustrated in the design table illustrated in Table 1.

Among the different equating methods, we will consider two widely used nonlinear observed score equating methods for the NEAT design: poststratification equating (PSE) and chain equipercentile equating (CE). The method known as *frequency estimation* is a type of PSE method. We will also consider the item-response-theory observed-score equating (IRT OSE) method. The IRT OSE is an IRT-based method that is comparable to the PSE and CE methods and is based on a theoretical model for the observed and missing data.

Research shows that the different equating methods may provide substantially different results under some conditions. Hence, it is important to compare these three equating methods and determine the most accurate one.

Table 1

The Design Table for the Nonequivalent Groups With Anchor Test (NEAT) Design

	X	*A*	*Y*
P	✓	✓	
Q		✓	✓

Note. A ✓-mark indicates that the test is given to a sample from the population to the left. *X* is not given to examinees from population *Q* while *Y* is not given to examinees from population *P*.

As Table 1 shows, the NEAT design involves missing data (i.e., data that are *missing by design* and not missing due to the various problems of individual respondents skipping questions or other types of testing realities). All data on X are missing in Q and all data on Y are missing in P**.** Each of the three equating methods considered here involves distinct assumptions about the data missing in the NEAT design. These missing data assumptions are not directly testable in practical equating situations. However, it is important to understand the missing data assumptions of the equating methods; we believe this understanding will improve the equating research and provide a better understanding of the methods. As we will show, each method will be optimal when its missing data assumptions are correct and will be less so when different assumptions are correct.

Though studies have compared equating performance of the PSE and CE methods under the NEAT design, no study examines the missing data assumptions of all three methods and discusses their implications on the comparison between the methods. Our work follows Holland, von Davier, Sinharay and Han (2006) and Holland, Sinharay, von Davier, and Han (2008), who described the missing data assumptions of the PSE and CE methods.

Several authors have given theoretical comparisons between two or more of these three methods. Harris and Kolen (1990) suggested that the PSE and the IRT OSE methods have superior theoretical underpinnings compared to the CE method while the latter method requires the least amount of computation. von Davier, Holland, and Thayer (2004a) provided some theoretical analyses of the PSE and CE methods and showed that they may both be interpreted as examples of OSE methods and that they can provide identical results under special conditions.

Several studies have compared the PSE and CE methods using real and simulated data sets. Sinharay (2008) provided a summary of these studies. Marco, Petersen, and Stewart (1983) and Livingston, Dorans, and Wright (1990) compared the PSE and CE methods (and several others) using data from the SAT®-I examination. Harris and Kolen (1990) compared the PSE and CE methods using data from a certification test. Sinharay and Holland (2007) compared the PSE and CE methods for several simulated data sets and a data set from a licensure test. Wang, Lee, Brennan, and Kolen (2006) compared the PSE and CE methods for several simulated data sets. The studies found that the CE method performed better than the PSE method when the two examinee groups, P and Q, differed substantially on the anchor test whereas the PSE method performed better than the CE method when the two examinee groups did not differ much. No study known to us compared the IRT OSE method to the PSE and CE methods, though Kolen and

Brennan (2004) recommended the operational use of the IRT OSE method, and the computations for the IRT OSE method did not take long with modern computers.

It can be argued that none of the above equating studies performed a complete and fair comparison of the PSE and CE methods. The theoretical comparisons do not suggest how the methods compare in a real testing environment. In the studies using real data sets, a lack of a clear criterion (or true equating function) for evaluating equating bias (or systematic error in equating) generally exists, as pointed out by Wang et al. (2006); a rare exception is the real data example in Sinharay and Holland (2007) where a clear criterion was present. The simulation studies by Wang et al. and Sinharay and Holland were well-designed and extensive, but both of these generated data using an IRT model and set up the criterion using an IRT procedure. It is possible that by generating data using an IRT model, they gave an undue advantage or disadvantage to one of the methods as expressed in the following quote from Wang et al.:

> …it is virtually impossible in a simulation to have a criterion that does not advantage/disadvantage one or more methods to some degree. In this particular simulation, it is possible that the criterion disadvantages the frequency estimation to some unknown extent. Specifically, there is nothing in the procedure used to establish the criterion that accommodates or reflects the assumption that, conditional on anchor test scores, Form X scores in the two populations are the same; similarly for Form *Y* scores. One could conceive of constructing the criterion in such a manner that these conditional distributions are indeed equal. If that were done, it seems reasonable to speculate that frequency estimation might have considerable less bias than exhibited in this study. (p. 15)

The above literature review shows a need to examine the missing data assumptions of the different methods and to understand their implications for any comparison of the three methods. Hence, the first objective of our project was to clearly state the missing data assumptions of the three equating methods in comparable ways and use them to simulate the missing data in order to give a deeper understanding of their effect on the choice of an equating criterion.

The literature review also shows a need to perform a fair comparison of the three methods. Our second goal was to perform such a comparison. To accomplish that, we filled in the missing data in different ways that reflected the missing data assumptions of each of the three methods.[1] When a particular missing data assumption held (that is, the missing data were filled in a way that reflected the missing data assumptions of a method), there was a corresponding criterion equating

function to which the three methods could be compared. Each method was shown to be optimal when its missing data assumptions held. In addition, the sensitivity of each method to the other missing data assumptions was also examined. This comparison constituted a study of the robustness of the method to missing data assumptions. The collection of these results provided us with a more complete comparison of the three methods than had been heretofore available.

The rest of this paper is organized as follows: Section 2 provides some background information, especially the concept of the target population and the true equating function in the NEAT design—these ideas will be relevant for the later sections. In Section 3, we describe the missing data assumptions for each of the three equating methods and their implications for the true equating function. Section 4 outlines our procedure to provide a fair comparison among the three equating methods in anticipation of the next three sections. Sections 5 to 7 discuss three examples based on real data: two teachers' licensing test examples and an admissions test example. Section 8 provides discussion and conclusions.

2. Background Information

Target Population in Equating and the Missing Data Assumptions of the Nonequivalent Groups With Anchor Test (NEAT) Design

We take the position that a given *linking function* may be interpreted as an OSE function if and only if (a) there is a specific target/synthetic population, T, on which the cumulative distribution functions (cdfs), $F_T(x) = \mathrm{P}\{X \leq x \mid T\}$ and $G_T(y) = \mathrm{P}\{Y \leq y \mid T\}$, are defined and (b) these cdfs are such that the linking function in question has the equipercentile form, that is, it is equal to the function,

$$y = G_T^{-1}(F_T(x)). \qquad (1)$$

Without this interpretation of OSE methods, the various linking functions for the NEAT design are merely procedures that have some intuitive plausibility but no particular relationship to the OSE methods that had been developed for the simpler designs (e.g., the more transparent single group or equivalent groups designs) which do make use of (1), and where T is the obvious population from which the data are sampled. Thus, to implement our approach, we need to define an appropriate synthetic population for the NEAT design and to specify appropriate cdfs, $F_T(x)$ and $G_T(y)$, for X and Y so that (1) holds for the linking functions in question.

A basic problem with the NEAT design is that the two populations, P and Q, can, in principle, be very different. This difference makes the definition of the target population more complex than it is for the simpler designs. Braun and Holland (1982) defined the synthetic population as the weighted mixture,

$$T = w\boldsymbol{P} + (1 - w)Q. \tag{2}$$

They interpreted (2) as a prescription for computing moments and distributions of variables over T. Their prescription is to first compute the quantities on P and on Q separately and then to combine these results with the weights w and $1 - w$. For example, to compute the mean of A over T, first compute the mean of A over P and Q separately, and then weight the two means by w and $1 - w$, respectively, to get the mean over T, for example,

$$\mathrm{E}(A \mid T) = w\mathrm{E}(A \mid P) + (1 - w)\mathrm{E}(A \mid Q). \tag{3}$$

Similarly, probabilities for A over T are computed as

$$\mathrm{P}\{A = a \mid T\} = w\,\mathrm{P}\{A = a \mid P\} + (1 - w)\,\mathrm{P}\{A = a \mid Q\}. \tag{4}$$

The total sample is obtained as a special case of Braun and Holland's (1982) synthetic population by choosing w as

$$w = N_{\boldsymbol{P}}/(N_{\boldsymbol{P}} + N_{\boldsymbol{Q}}), \tag{5}$$

where $N_{\boldsymbol{P}}$ and $N_{\boldsymbol{Q}}$ denote the sample sizes from P and Q, respectively. Thus, the total sample, $P + Q$, is the synthetic population that weights P and Q proportionally to their sample sizes. We will use this choice of w in this paper. However, other choices of w might be of interest, for example $w = 1$ (equate on the new population, P) or $w = 0$ (equate on the old population, Q). For this reason, w is a parameter of the definition of any particular synthetic population, T.

The prescription for computing moments and probabilities in (3) and (4) is straightforward when applied to A. However, when it is applied to X or Y, it requires finding values for the distributions of X or Y over populations for which *they are always missing*. That is, there is a need to specify distributions (or at least some moments) of X in Q and Y in P.

Thus, the use of the synthetic population in (2) and the definition of observed score equating in (1) require attention to missing data assumptions that are often tacit and unstated

when a equating function is simply described as a procedure for finding the *y*-values that are equated to an *x*-value. The important feature of these missing data assumptions is that they allow the distributions or moments of *X* in *Q* and *Y* in *P* to be computed. These moments or distributions then allow the corresponding moments or distributions for *X* and for *Y* to be computed for any synthetic population given by (2) using (3) or (4) with *X* and *Y* replacing *A*. Once the distribution or moments of *X* and *Y* on *T* are available, the NEAT design effectively becomes a single group design over the specified synthetic population ***T,*** and the OSE methods defined by (1) may be used.

The True Equating Functions for the Nonequivalent Groups With Anchors Test (NEAT) Design

In equating studies, it is important to have a criterion to compare to the equating functions that are estimated from the data. We shall call the criterion the *true equating function* in this paper. It is usually the case that *true* is synonymous with *population* in the sample/population distinction (i.e., the true mean refers to the population mean as opposed to the sample mean). In equating studies it is rare to have a population in hand, because the data are samples that may be large but still exhibit the variability typical of such samples. This study is no exception, and we recognize that in using the term *true equating function,* we are glossing over the difference between a large sample and the population. However, for the NEAT design an added complexity is present. It is impossible to define a true equating function without regard to the assumptions that are made about the missing data. Each choice of missing data assumption leads to a possibly different true equating function. Other factors influence the proper choice of the true equating function for a given estimated equating function as well. We will discuss these below.

What is the form, in general, of the true equating function for the NEAT design? First, observe that if the missing data of the NEAT design displayed in Table 1 were not missing, then any synthetic population, $T = w\boldsymbol{P} + (1 - w)Q$, would simply be the single population formed by weighting the cases from *P* and *Q* appropriately. For example, if *w* were given by (5), then the resulting equating design is the single group design with $P + Q$ as the population. The design table for the single group design is displayed in Table 2.

The natural way to equate *X* to *Y* in this single group design is to form the two cdfs, $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$, from the data for *X* and *Y* over *T* and then to use formula (1) to equate scores on *X* to those on *Y*.

Table 2

The Design Table for the Single Group Design

	X	*Y*
T	✓	✓

Note. A ✓-mark indicates that the test is given to a sample from the population to the left.

In view of this observation, we argue that an appropriate way to define the true equating function for the NEAT design involves the following three steps:

1. Make specific missing-data assumptions and use them to deduce the implied distributions of X in Q and Y in P, for example, the missing data marginals,

$$\mathrm{P}\{X = x \mid Q\} \text{ and } \mathrm{P}\{Y = y \mid P\}.$$

2. Combine the observed distributions of X in P and Y in Q with the implied missing data marginals from the previous step to derive the corresponding cdfs of X and Y on T as

$$F_T(x) = \mathrm{P}\{X \le x \mid T\} = w\,\mathrm{P}\{X \le x \mid P\} + (1 - w)\,\mathrm{P}\{X \le x \mid Q\} \text{ and} \quad (6)$$

$$G_T(y) = \mathrm{P}\{Y \le y \mid T\} = w\,\mathrm{P}\{Y \le y \mid P\} + (1 - w)\,\mathrm{P}\{Y \le y \mid Q\}. \quad (7)$$

3. Compute the equipercentile equating function as $y = G_T^{-1}(F_T(x))$ using the cdfs obtained from (6) and (7).

The result of Step 3 is the true equating function that corresponds to the missing data assumptions specified in Step 1 for the synthetic population T in Step 2. The missing data assumptions enter only through their implications for the missing data marginals. We used this missing data marginal approach to compute the various true equating functions in this study.

This approach should be compared to an alternative true equating function calculation that may seem more natural to some readers. In this alternative, using the three equating methods, the missing data for each examinee are filled in so that everyone in both P and Q has an X, A, and Y score (two real scores and one imputed from the missing data assumptions). The two data sets are

weighted appropriately and pooled to form $T = w\boldsymbol{P} + (1 - w)Q$. The cdfs $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ are formed from it and the equipercentile equating function is computed using (1). We call this approach *filling-in the missing data*. We used both the missing data marginal approach and the filling-in the missing data approach in our data examples (to be discussed later), but found only small differences between the resulting true equating functions. Hence, for the most part, we report only the results using the missing data marginals method in our examples, but we show the results of both approaches in our first data example.

Our approach of defining a true equating function for the NEAT design using the missing data marginals leaves open three important decisions.

First, which missing data assumptions should be used to obtain the missing data marginals $\mathrm{P}\{X = x \mid Q\}$ and $\mathrm{P}\{Y = y \mid P\}$? Different missing data assumptions will lead to different values for these probabilities and then to different true equating functions.

Second, while the missing data assumptions will give us values for the missing data marginals, $\mathrm{P}\{X = x \mid Q\}$ and $\mathrm{P}\{Y = y \mid P\}$, we also need values for the observed data marginals, $\mathrm{P}\{X = x \mid P\}$ and $\mathrm{P}\{Y = y \mid Q\}$ to use (6) and (7). How should the observed data for X in P and Y in Q be used to provide this information?

Third, how should the cdfs, $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$, which are used in equation (1) to produce the true equipercentile function, be calculated?

Regarding the second question on the use of the observed data for X in P and Y in Q, we chose to presmooth, that is, fit a loglinear model (e.g., Holland & Thayer, 2000) to the observed frequencies. We used the presmoothed sample proportions as estimates of the population probabilities.

Regarding the third question on the calculation of the cdfs $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$, we used the traditional method of continuizing the discrete score distributions that linearly interpolates the discrete cdfs to obtain continuous cdfs (e.g., Kolen & Brennan, 2004).[2]

Clarifying the role of the missing data assumptions, which was asked in the first question, is the least routine of the three and we address that in the next section.

3. The Missing Data Assumptions for Poststratification Equating (PSE), Chain Equipercentile Equating (CE) and Item-Response-Theory Observed-Score Equating (IRT OSE)

We provide a description of the missing data assumptions of the three equating methods below. We do this in two ways for each set of assumptions. First, in the missing data marginals

approach, we give expressions for the missing data marginals, $P\{X = x \mid Q\}$ and $P\{Y = y \mid P\}$. Second, in the filling-in the missing data approach, we show how to fill in the missing data values for each examinee in the NEAT design so that the missing data assumptions for each of the methods—PSE, CE, or IRT OSE—are satisfied exactly.

Strictly speaking, it is not necessary to fill in the missing data for each examinee in order to obtain the true equating function. All that is required are the values of the missing data marginals, $P\{X = x \mid Q\}$ and $P\{Y = y \mid P\}$. However, filling in the missing data in a data set in different ways allows us to study how the different features of the data are affected by the different missing data assumptions. For example, do the different ways of filling in the missing data lead to significantly different correlations between the scores on X and those on A over the combined single population, T? This total-test to anchor-test correlation is known to influence the quality of an equating.

Missing Data Assumptions of Poststratification Equating (PSE)

The assumptions of the PSE method are that the conditional distributions of X and Y given A are the same in P and in Q. More formally,

$$P\{X = x \mid A = a, Q\} = P\{X = x \mid A = a, P\} \text{ and}$$

$$P\{Y = y \mid A = a, P\} = P\{Y = y \mid A = a, Q\}. \tag{8}$$

We first discuss how (8) can be used to specify the missing data marginals and then discuss how to create score triples by filling in the missing values in the NEAT design so that the PSE assumptions are satisfied.

The Missing Data Marginals and the True Equating Function Implied by Poststratification Equating (PSE)

To compute the missing data marginal $P\{X = x \mid Q\}$ implied by the PSE assumptions, first use the (X, A) data from P to estimate the conditional distribution $P\{X = x \mid A = a, P\}$ and then use the PSE assumption (8) to treat this as an estimate of $P\{X = x \mid A = a, Q\}$. Finally, use the standard formula for obtaining marginal probabilities from joint probabilities to obtain $P\{X = x \mid Q\}$, for example,

$$\mathrm{P}\{X = x \mid Q\} = \sum_a \mathrm{P}\{X = x \mid A = \mathrm{a}, P\}\ \mathrm{P}\{A = a \mid Q\}. \qquad (9)$$

Similarly, to compute the missing data marginal $\mathrm{P}\{Y = y \mid P\}$ from the (Y, A) data from Q, apply the PSE assumption in (8) to get

$$\mathrm{P}\{Y = y \mid P\} = \sum_a \mathrm{P}\{Y = y \mid A = \mathrm{a}, Q\}\ \mathrm{P}\{A = a \mid P\}. \qquad (10)$$

Equations (9) and (10) are the key ingredients to the PSE method.

To form the probability estimates needed to implement (9) and (10), we used bivariate presmoothing (Holland & Thayer, 2000). Denote these presmoothed bivariate distributions by

$$pxa = \mathrm{P}\{X = x, A = a \mid P\} \text{ and } qya = \mathrm{P}\{Y = y, A = a \mid Q\}, \qquad (11)$$

and the implied presmoothed marginal distributions of A in P and Q by

$$ha_{\boldsymbol{P}} = \mathrm{P}\{A = a \mid P\} \text{ and } ha_{\boldsymbol{Q}} = \mathrm{P}\{A = a \mid Q\}. \qquad (12)$$

The estimates of the conditional probabilities that are needed in (9) and (10) are given by the ratios:

$$\mathrm{P}\{X = x \mid A = \mathrm{a}, P\} = pxa/ha_{\boldsymbol{P}} \text{ and } \mathrm{P}\{Y = y \mid A = \mathrm{a}, Q\} = qya/ha_{\boldsymbol{Q}}. \qquad (13)$$

Using (9), the missing data marginal for X is obtained by weighting $p_{xa}/h_{a\boldsymbol{P}}$ by $h_{a\boldsymbol{Q}}$ and summing over a. The missing data marginal for Y is obtained in a similar way from $q_{ya}/h_{a\boldsymbol{Q}}$ weighted by $h_{a\boldsymbol{P}}$.

After the missing data marginals, $\mathrm{P}\{X = x \mid Q\}$ and $\mathrm{P}\{Y = y \mid P\}$, are obtained, there are two ways of using them to obtain the cdfs, $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$. The first, which we employ here, is to continuize the four discrete distributions, $\mathrm{P}\{X = x \mid P\}$, $\mathrm{P}\{X = x \mid Q\}$, $\mathrm{P}\{Y = y \mid P\}$, and $\mathrm{P}\{Y = y \mid Q\}$[3] to form the four cdfs, $F_{\boldsymbol{P}}(x)$, $F_{\boldsymbol{Q}}(x)$, $G_{\boldsymbol{P}}(y)$, and $G_{\boldsymbol{Q}}(y)$. These four cdfs may then be combined as in (6) and (7) to form $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ which are then used in (1) to obtain the PSE true equating function. In all our calculations of equating functions we used the iterative Illinois method (e.g., Thisted, 1988, p. 171) to find the inverse required in (1).

However, the kernel equating version of PSE (von Davier, Holland, & Thayer, 2004b) does the calculation of $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ in a slightly different way. They propose to first combine $\mathrm{P}\{X = x \mid P\}$ and $\mathrm{P}\{X = x \mid Q\}$ via the equation analogous to (4) to obtain $\mathrm{P}\{X = x \mid T\}$ and combine $\mathrm{P}\{Y =$

$y \mid P\}$ and $\mathrm{P}\{Y = y \mid Q\}$ to obtain $\mathrm{P}\{Y = y \mid T\}$. Then, $\mathrm{P}\{X = x \mid T\}$ and $\mathrm{P}\{Y = y \mid T\}$ are continuized to get $F_T(x)$ and $G_T(y)$ and then (1) is applied to obtain the estimated equating function for PSE. We use this approach to compute the estimated equating function for the PSE method. It uses only two continuizations, while the one we used to compute the true equating functions required four. However, whether two continuizations were used or four mattered little for our data examples; as a result, for each data example, the true equating function under the PSE assumptions was virtually identical to the estimated equating function from the PSE method.

Filling in the Missing Data in Poststratification Equating (PSE) Score Triples

The following approach may be used to create score triples (X, A, Y) that satisfy the PSE assumptions. For each (X, A) pair from P, take the A score and draw a random Y value from the presmoothed conditional distribution of Y given that A value from the Q sample, $\mathrm{P}\{Y = y \mid A = \mathrm{a}, Q\} = q_{ya}/h_{a\mathbf{Q}}$. The result is an ($X$, A, Y) score triple for each examinee in P that satisfies the PSE assumptions where X and A are the observed data and Y is a filled-in missing data value under the PSE missing data assumptions. Similarly, use the same techniques to fill in an X-value to obtain a score triple for each examinee in Q.

Note that, by construction, this method makes the distribution of Y conditionally independent of X given A in both P and Q.

Once the score triples are created, the design table is now filled in with X, Y, and A scores for every examinee in P and Q. These filled-in data may be used as described in Section 2 to compute the true equating function for PSE. It is also possible to study features of the filled-in values and their relations to the observed data.

The Missing Data Assumptions of Chained Equipercentile Equating (CE)

The assumptions of CE do not directly involve the discrete score distributions in the way that the assumptions of PSE do. Rather, they may be viewed as assumptions about two equipercentile functions similar to those in (1). For our purposes, the CE missing data assumptions are that the equipercentile functions equating X to A are the same in P and Q; the same is true for the equipercentile links from A to Y in P and Q.

So far we have used F for the cdf of X, G for the cdf of Y and now we use H for the cdf of A, with appropriate subscripts to indicate the populations to which the cdfs refer. We assume that

all the discrete distributions have been continuized so that all the relevant cdfs are continuous and strictly increasing.

The assumption that the equipercentile link from X to A is the same in P and Q may be formalized as

$$H_Q^{-1}(F_Q(x)) = H_P^{-1}(F_P(x)). \quad (14)$$

Similarly, the assumption that the equipercentile link from A to Y is the same for P and Q is formalized as

$$G_Q^{-1}(H_Q(a)) = G_P^{-1}(H_P(a)). \quad (15)$$

Because (14) and (15) involve the cdfs of the missing data, for example, $F_Q(x)$, and $G_P(y)$, they qualify as missing data assumptions, and we now explore them further.

The Missing Data Marginals and the True Equating Function Implied by Chained Equipercentile Equating (CE)

Solving for $F_Q(x)$ in (14) yields

$$F_Q(x) = H_Q(H_P^{-1}(F_P(x))). \quad (16)$$

This equation connects the cdf of the missing data to the cdfs of data that are observed. Equation (16) is the CE missing data assumption for X and A. The expression for $F_Q(x)$ in (16) is a continuous cdf. We interpret it as the continuized cdf of the discrete missing data marginal for X in Q. Thus, in the case of CE we obtain the continuization of the missing data marginal directly rather than the missing data marginal itself. We may then combine $F_Q(x)$ with $F_P(x)$, the continuized cdf obtained from the observed distribution of X in P, to obtain

$$F_T(x) = wF_P(x) + (1 - w)F_Q(x). \quad (17)$$

In a similar way we may solve (15) to obtain the continuization of the missing data marginal of Y in P as

$$G_P(y) = H_P(H_Q^{-1}(G_Q(y))). \quad (18)$$

Similarly, we may then compute $G_T(y)$ from $G_P(y)$ and $G_Q(y)$ to obtain

$$G_{\boldsymbol{T}}(y) = wG_{\boldsymbol{P}}(y) + (1 - w)G_{\boldsymbol{Q}}(y). \quad (19)$$

The true equating function is obtained by applying (1) to $F_{\boldsymbol{T}}$ and $G_{\boldsymbol{T}}$ from (17) and (19).

The usual form of the estimated CE equipercentile function is given by the following chain of cdfs and their inverses (e.g., von Davier et al., 2004b, p. 37),

$$G_{\boldsymbol{Q}}^{-1}(H_{\boldsymbol{Q}}(H_{\boldsymbol{P}}^{-1}(F_{\boldsymbol{P}}(x)))). \quad (20)$$

In (20), $H_{\boldsymbol{P}}^{-1}(F_{\boldsymbol{P}}(x))$ is the X to A link on P and $G_{\boldsymbol{Q}}^{-1}(H_{\boldsymbol{Q}}(a))$ is the A to Y link on Q and the CE equipercentile function consists of the functional composition of these two links. The question naturally arises as to whether the equipercentile function we have specified in Section 2 as the true equating function for CE agrees with formula (20). Under the CE missing data assumptions, it does, and this result is summarized in Theorem 1, below.

Theorem 1: If $F_{\boldsymbol{Q}}(x) = H_{\boldsymbol{Q}}(H_{\boldsymbol{P}}^{-1}(F_{\boldsymbol{P}}(x)))$ and $G_{\boldsymbol{P}}(y) = H_{\boldsymbol{P}}(H_{\boldsymbol{Q}}^{-1}(G_{\boldsymbol{Q}}(y)))$, that is, if the CE missing data assumptions hold, then for any target population of the form $T = w\boldsymbol{P} + (1 - w)Q$ and $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(x)$ defined in (17) and (19), the following equation holds,

$$G_{\boldsymbol{T}}^{-1}(F_{\boldsymbol{T}}(x)) = G_{\boldsymbol{Q}}^{-1}(H_{\boldsymbol{Q}}(H_{\boldsymbol{P}}^{-1}(F_{\boldsymbol{P}}(x)))).$$

Because the right-hand side above is the usual form of the estimated CE equipercentile function (i.e., equation 20), Theorem 1 shows that the true equating function obtained from the CE missing data assumptions exactly equals the estimated CE equipercentile function. The proof of Theorem 1 is left to Appendix B. The results from our real data examples are coherent with Theorem 1.

Filling in the Missing Data in the Chained Equipercentile Equating (CE) Score Triples

The score pairs (X, A) in P may be augmented to score triples (X, A, Y) that satisfy the CE assumptions. The idea is to find a conditional distribution, $\mathrm{P}^*\{Y = y \mid A = \mathrm{a}, P\}$, that reflects the CE assumptions and then, for each examinee in P, to use the value of A to draw a Y-value from this conditional distribution to form a score triple (X, A, Y). This way is similar our suggestion for simulating score triples that satisfy the PSE missing data assumptions, but the CE case requires a more complicated approach to find the appropriate conditional distributions. We describe the process in several steps.

1. Form the presmoothed marginal discrete distributions for X in P, A in P, A in Q and Y in Q by summing the rows and columns of $\{p_{xa}\}$ and $\{q_{ya}\}$ defined in (11) and denote these marginals respectively as $\{f_{x\boldsymbol{P}}\}$, $\{h_{a\boldsymbol{P}}\}$, $\{h_{a\boldsymbol{Q}}\}$, and $\{g_{y\boldsymbol{Q}}\}$.

2. Continuize these four discrete distributions to obtain the continuous cdfs, $F_{\boldsymbol{P}}(x)$, $H_{\boldsymbol{P}}(a)$, $H_{\boldsymbol{Q}}(a)$, and $G_{\boldsymbol{Q}}(y)$. Up to this point, everything is part of computing the CE equipercentile function. Now we use the missing data marginals implied by the CE method to get to the missing data.

3. Compute the missing data cdfs, $F_{\boldsymbol{Q}}(x) = H_{\boldsymbol{Q}}(H_{\boldsymbol{P}}^{-1}(F_{\boldsymbol{P}}(x)))$ and $G_{\boldsymbol{P}}(y) =$

$$H_{\boldsymbol{P}}(H_{\boldsymbol{Q}}^{-1}(G_{\boldsymbol{Q}}(y))), \text{ as in (16) and (18).}$$

4. Discretize $F_{\boldsymbol{Q}}(x)$ to obtain a version of the discrete missing data marginal that would continuize to the cdf $F_{\boldsymbol{Q}}(x)$. We propose using a method described in Holland et al. (2008, 2006). They suggested defining the half-score points as $x_j^* = (x_j + x_{j+1})/2$, for $j =$ 1 to $J - 1$, and then defining the discrete missing data marginal of X in Q as

$$\begin{aligned} &\mathrm{P}\{X = x_j \mid Q\} = F_{\boldsymbol{Q}}(x_j^*) - F_{\boldsymbol{Q}}(x_{j-1}^*), \text{ for } j = 2, 3, \ldots, J-1, \\ &\mathrm{P}\{X = x_1 \mid Q\} = F_{\boldsymbol{Q}}(x_1^*), \text{ and } \mathrm{P}\{X = x_J \mid Q\} = 1 - F_{\boldsymbol{Q}}(x_{J-1}^*). \end{aligned} \tag{21}$$

Equation (21) implies that all values of x in an interval bracketed by the half-score points x_{j-1}^* and x_j^* are rounded to the raw score x_j with appropriate adjustments for plus and minus infinity. This method is a very natural way to discretize $F_{\boldsymbol{Q}}(x)$. $G_{\boldsymbol{P}}(y)$ is discretized in a similar way. These discretized distributions are the CE missing data marginals,

$$fx_{\boldsymbol{Q}} = \mathrm{P}\{X = x \mid Q\} \text{ and } gy_{\boldsymbol{P}} = \mathrm{P}\{Y = y \mid P\}, \tag{22}$$

for X in Q and Y in P.

The discrete distributions in (22) are compatible with the CE assumptions in the sense that if we continuize them we will get cdfs that are either very close to or identical to $F_{\boldsymbol{Q}}(x)$ and $G_{\boldsymbol{P}}(y)$ given in (16) and (18). These missing data marginals therefore reflect the CE assumptions in a very specific way.

5. Use *raking* (also called *iterative proportional fitting* or the *Deming-Stephan method*), described in Bishop, Fienberg, & Holland (1975), to transform the presmoothed joint distribution $\{p_{xa}\}$ so that it has the missing data marginal of X in Q, $\{f_{x\mathbf{Q}}\}$, from (22) as its X-marginal, and the marginal of A in Q, $\{h_{a\mathbf{Q}}\}$, from (12) as its A-marginal. The process of raking is described in Appendix A.

Call the result of this raking

$$kxa_{\mathbf{Q}} = \mathrm{P}\{X = x, A = a \mid Q\}. \tag{23}$$

The point of the raking is to create a joint distribution for (X, A) that has the marginal distributions of X and A in Q but the odds ratios or cross-product ratios (that measure the association/correlation between X and A) of the presmoothed joint distribution of (X, A) in P.

6. Form the desired conditional distribution by forming the appropriate ratios,

$$\mathrm{P}^*\{X = x \mid A = a, Q\} = kxa_{\mathbf{Q}} / ha_{\mathbf{Q}}. \tag{24}$$

Similarly, rake $\{q_{ya}\}$ using $\{g_{y\mathbf{P}}\}$ from (22) and $\{h_{a\mathbf{P}}\}$ from (12). Call the result of this raking

$$k_{ya\mathbf{P}} = \mathrm{P}\{Y = y, A = a \mid P\}. \tag{25}$$

The result of the second raking is a joint distribution that has the marginal distributions $\{g_{y\mathbf{P}}\}$ and $\{h_{a\mathbf{P}}\}$ that are appropriate for P and has the cross-product ratios of the presmoothed joint distribution of (Y, A) in Q. From this joint distribution form the desired conditional distribution by

$$\mathrm{P}^*\{Y = y \mid A = a, P\} = k_{ya\mathbf{P}} / h_{a\mathbf{P}}. \tag{26}$$

Note that because of the raking procedure, the A marginals formed from the bivariate distributions in (23) and (25) are the appropriate divisors so that the results in (24) and (26) are *bona fide* conditional probabilities. We used the symbol P* in (24) and (26) to emphasize that these conditional distributions need not be the same as those used in PSE.

For each A value for a score pair (Y, A) in Q, use P* in (24) to draw a random X value and thereby form a score triple (X, A, Y) for examinees in Q. Analogously, draw a random Y value from (26) to form score triples for examinees in P. Note that, as in the case of the PSE triples, by construction, X and Y are conditionally independent given A, in both P and Q.

Missing Data Assumptions of Item-Response-Theory Observed-Score Equating (IRT OSE)

The missing data assumptions of IRT OSE include the assumption that the IRT model is adequate for the observed data in addition to the specific missing data assumptions indicated below. We used the two-parameter logistic (2PL) model throughout our analyses. Holland (1990) showed that enough information is present to estimate at most two parameters per item in applications of IRT, and Haberman (2006) demonstrated that the 2PL model describes real test data as well as the three-parameter logistic (3PL) model. In addition to providing another source of missing data assumptions, the IRT model also serves as a form of presmoothing for all the relevant discrete distributions and is an alternative to fitting loglinear models to the score distributions.

The Item Response Theory (IRT) Model

The discussion for IRT OSE uses with test scores X, Y, and so on, but it is important to remember that estimation of the IRT model requires item level data.

The item response functions may be summed appropriately to determine the score response functions, $P_X(x\,|\,\theta, \boldsymbol{\beta}_{XP})$, $P_Y(y\,|\,\theta, \boldsymbol{\beta}_{YQ})$, $P_A(a\,|\,\theta, \boldsymbol{\beta}_{AP})$, and $P_A(a\,|\,\theta, \boldsymbol{\beta}_{AQ})$, which are defined by

$$P_X(x\,|\,\theta, \boldsymbol{\beta}_{XP}) = \mathrm{P}\{X = x \mid \theta, P\},\ P_Y(y\,|\,\theta, \boldsymbol{\beta}_{YQ}) = \mathrm{P}\{Y = y \mid \theta, Q\},$$

$$P_A(a\,|\,\theta, \boldsymbol{\beta}_{AP}) = \mathrm{P}\{A = a \mid \theta, P\}, \text{ and } P_A(a\,|\,\theta, \boldsymbol{\beta}_{AQ}) = \mathrm{P}\{A = a \mid \theta, Q\}. \tag{27}$$

To compute the score response functions, for example, $\mathrm{P}\{X = x \mid \theta, P\}$, from the item response functions, we use the recursive formula of Lord and Wingersky (1984). The parameters such as $\boldsymbol{\beta}_{XP}$ in (27) are the vectors of relevant item parameters for the items in each score.

The subscripts P and Q in (27) indicate that the $\boldsymbol{\beta}$-parameters might differ in P and Q, that is, in general, they might show differential item functioning (DIF). The only items for which the presence of DIF can be tested are those in A, which have item responses for both populations. An important assumption for the NEAT design is that the A items do not show DIF. Hence, for the rest of our discussion we will assume that

$$\boldsymbol{\beta}_{AP} = \boldsymbol{\beta}_{AQ} = \boldsymbol{\beta}A. \tag{28}$$

Thus, the missing data assumptions for IRT OSE are that the items in X and Y (in addition to those in A) do not exhibit DIF across P and Q, so that the score response functions for X or Y given θ are the same for P and Q. To indicate this, henceforth, we will drop the subscripts P and Q in (27) as we did in (28).

In addition to item response functions, the IRT model for X, Y, and A requires θ distributions for both P and Q. Because these populations can be different in the NEAT design, these θ distributions should be allowed to be different in the estimation process. Denote the densities of these two θ distributions by $k_P(\theta)$ and $k_Q(\theta)$. For many problems, it is satisfactory to assume that $k_P(\theta)$ and $k_Q(\theta)$ are both Gaussian with possibly different means and variances, that is

$$k_P(\theta) = \phi(\theta) \text{ and } k_Q(\theta) = (1/\sigma)\phi((\theta - \mu)/\sigma), \tag{29}$$

where ϕ denotes the standard normal density function. The usual location/scale indeterminacy of IRT forces one of the Gaussians densities to have mean zero and variance one. We assumed Gaussian ability distributions in our work.

Proper estimation of the IRT model for the NEAT design includes estimation of μ and σ in (29) as well as the item parameters, β_X, β_Y, and β_A in (27). To estimate the parameters of the 2PL model, we used a separate calibration (using marginal maximum likelihood) of the model for X and A in P and for A and Y in Q, assuming the standard normal θ distribution for each. We then used the Stocking-Lord method (see, e.g., Kolen & Brennan, 2004) to convert the 2PL model item parameter estimates for P and Q to the same scale and to obtain the corresponding estimates of μ and σ in (29).

Assuming that a satisfactory IRT model has been estimated for the NEAT design and the parameters have been converted to the same scale, any synthetic population of the form in (2) is handled by simply using the same weights for the two θ distributions, such as,

$$k_T(\theta) = w\, k_P(\theta) + (1 - w)\, k_Q(\theta). \tag{30}$$

In standard IRT OSE, the densities defined in (29) and (30) are used to integrate out θ in the conditional distributions in (27) to form the marginal distributions of X and Y over T. For example,

$$\mathrm{P}\{X = x \mid T\} = \int P_{\boldsymbol{X}}(x \mid \theta, \boldsymbol{\beta}_X)\, k_{\boldsymbol{T}}(\theta) d\theta, \text{ and}$$

$$\mathrm{P}\{Y = y \mid T\} = \int P_{\boldsymbol{Y}}(y \mid \theta, \boldsymbol{\beta}_Y)\, k_{\boldsymbol{T}}(\theta) d\theta. \tag{31}$$

The Missing Data Marginals and the True Equating Function for Item-Response-Theory Observed-Score Equating (IRT OSE)

By the same reasoning that leads to (31), the missing data marginals for IRT OSE are

$$\mathrm{P}\{X = x \mid Q\} = \int P_{\boldsymbol{X}}(x \mid \theta, \boldsymbol{\beta}_X)\, k_{\boldsymbol{Q}}(\theta) d\theta, \text{ and}$$

$$\mathrm{P}\{Y = y \mid P\} = \int P_{\boldsymbol{Y}}(y \mid \theta, \boldsymbol{\beta}_Y)\, k_{\boldsymbol{P}}(\theta) d\theta. \tag{32}$$

We used the Gauss-Hermite quadrature method to compute the above integrals. Because the quantities in the integral in (32) are all smooth functions, the resulting values for the IRT OSE missing data marginals are a form of presmoothing the raw score distributions and do not exhibit the noisy extra variability of raw frequencies.

After the missing data marginals, $\mathrm{P}\{X = x \mid Q\}$ and $\mathrm{P}\{Y = y \mid P\}$, are obtained as described above, they are continuized, then $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ are computed using (6) and (7), and the true equating function is obtained using (1).

Note that the computation of the IRT OSE estimated equating function proceeds in a similar manner to that for the IRT OSE true equating function by first continuizing the discrete distributions obtained in (31) to compute $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ and then by applying (1). Thus, the computation of the IRT OSE estimated equating function involves two continuizations whereas the computation of the IRT OSE true equating function requires four. We found negligible differences between the IRT OSE estimated equating function and IRT OSE true equating function in all three of the data examples discussed later.

Filling in the Missing Data in the Item-Response-Theory Observed-Score Equating (IRT OSE) Score Triples

The IRT model can be used to estimate the conditional distribution of Y given A and X over the population P as

$$\mathrm{P}\{Y = y \mid X = x, A = a, P\} = \int P_{Y}(y \mid \theta, \beta Y)\, k(\theta \mid X = x, A = a, P)\, d\theta, \qquad (33)$$

where the conditional density of θ given X and A over P, $k(\theta \mid X = x, A = a, P)$, is equal to

$$P_{X}(x \mid \theta, \beta X)\, P_{A}(a \mid \theta, \beta A)\, k_{P}(\theta) / [\int P_{X}(x \mid \tau, \beta X)\, P_{A}(a \mid \tau, \beta A)\, k_{P}(\tau) d\tau], \qquad (34)$$

if A is an external anchor and is equal to

$$P_{X'}(x' \mid \theta, \beta X)\, P_{A}(a \mid \theta, \beta A)\, k_{P}(\theta) / [\int P_{X'}(x' \mid \tau, \beta X)\, P_{A}(a \mid \tau, \beta A)\, k_{P}(\tau) d\tau], \qquad (35)$$

if A is an internal anchor, where $X=X'+A$ (that is, X can be partitioned into X' and A).

We computed the discrete conditional distribution in (33) by using Gauss-Hermite quadrature to perform the integration with respect to θ. To fill in the Y values to form a score triple for X, A, and Y in P, we made a random Y draw from the conditional distribution in (33) for each observed (X, A) pair in the data.

The same method may be applied to fill in X in (X, A, Y) for the Q sample. The conditional distribution of X given Y and A over Q is given by

$$\mathrm{P}\{X = x \mid Y = y\, , A = a, Q\} = \int P_{X}(x \mid \theta, \beta X)\, k(\theta \mid Y = y, A = a, Q)\, d\theta, \qquad (36)$$

where the conditional density of θ given Y and A over Q, $k(\theta \mid Y= y, A = a, Q)$, is equal to

$$P_{Y}(y \mid \theta, \beta Y)\, P_{A}(a \mid \theta, \beta A) k_{Q}(\theta) / [\int P_{Y}(y \mid \tau, \beta Y)\, P_{A}(a \mid \tau, \beta A)\, k_{Q}(\tau) d\tau], \qquad (37)$$

if A is an external anchor, and is equal to

$$P_{Y'}(y' \mid \theta, \beta Y)\, P_{A}(a \mid \theta, \beta A) k_{Q}(\theta) / [\int P_{Y'}(y' \mid \tau, \beta Y)\, P_{A}(a \mid \tau, \beta A)\, k_{Q}(\tau) d\tau], \qquad (38)$$

if A is an internal anchor, where $Y=Y'+A$.

Note that unlike either the PSE or CE cases described above, for the IRT OSE filled-in missing data, X and Y need not be conditionally independent given A over P or Q. This is because both X and A can influence the probability of Y in P and both Y and A can influence the probability of X in Q. It is possible to use a method similar to that described above but that uses only A to fill

in the missing data under the IRT model. This method will make X and Y conditionally independent given A, however, it does not make full use of the IRT model.

4. Our Measures of Robustness Against Missing Data Assumptions

We apply our analysis of the missing data assumptions for PSE, CE, and IRT OSE to three different data sets from operational tests using the NEAT design in the next three sections. Here we outline our analysis strategies, pulling together the ideas described in the previous sections.

One of our goals was to show how the missing data assumptions for a method lead to a true equating function that favors one of the methods and not the others. Any true equating function is virtually identical to the estimated equating function obtained using the same method and different from the estimated equating functions obtained using the other two methods. An equating method is robust against the other missing data assumptions if it performs reasonably well under various plausible missing data assumptions.

To help express the idea of robustness, for each of our three data sets, we give a table with three rows and three or four columns. The rows correspond to the three estimated equating functions—CE, PSE, or IRT OSE—while the first three columns correspond to the three missing data assumptions that lead to different true equating functions. The first two of our data sets are very special because the data that are usually missing in the NEAT design are not missing in these data sets. For these two cases the table has a fourth column (real data) that corresponds to the true equating function obtained using the real data as follows:

- We formed the bivariate distribution of X and Y over $P + Q$ and presmoothed it using appropriate loglinear models.
- We continuized the resulting marginal distributions of X and Y to obtain $F_{\boldsymbol{T}}$ and $G_{\boldsymbol{T}}$.
- Finally, $F_{\boldsymbol{T}}$ and $G_{\boldsymbol{T}}$ were combined as in (1) to obtain the true equating function for the real data case.

The cells of these tables contain various measures of the differences between the estimated equating function for that row and the true equating function for that column. Table 3 gives a schematic representation of these tables of differences.

Table 3

A Schematic Display of the Comparisons of the Estimated Equating Functions and the True Equating Functions Used in the Examples Sections

The estimated equating function	The assumptions behind the true equating function			
	CE	PSE	IRT OSE	Real data
CE	X	X	X	X
PSE	X	X	X	X
IRT OSE	X	X	X	X

Note. Each cell contains five discrepancy measures denoted by X. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

To form the tables illustrated by Table 3, we first used the data to calculate the estimated equating function for the three methods. We then calculated the true equating function that corresponded to each set of missing data assumptions (or the real data when these are available). The true equating function under each set of missing data assumptions was computed in two different ways based on the choice of the use of missing data marginals versus the use of the filled-in missing data to compute the true equating function.

Finally, we formed the differences between each estimated equating function and each true equating function. This resulted in either 9 or 12 sets of differences between the estimated and true equating functions that correspond to the cells of Table 3. The smaller the magnitude of the differences, the better the performance of the equating method is under the given missing data assumptions. To give numerical summaries of these function differences, we computed four discrepancy measures for each situation. These are as follows: (a) the root mean squared differences (*RMSD)* = $\sqrt{\sum_i \omega_i (e_i - T_i)^2}$, where ω_i denotes the weight and is proportional to the frequency of an *X*-score in the combined population *P*+*Q* when data on *X* in *Q* are available, or just in *P* when such data are not available, and e_i and T_i respectively denote the estimated and true equating functions at score-point *i*; (b) same as (a) but summed over only the lowest 25% of the examinees (i.e., those with *X*-scores below the 25th percentile of the frequency distribution of *X*);

(c) same as (a) but summed over only the middle 50% of the examinees (i.e., those with *X*-scores between the 25th and the 75th percentiles of the frequency distribution of *X*); and (d) same as (a) but summed over only the highest 25% of the examinees. In each cell of the tables that Table 3 illustrates, the four measures appear vertically in the above order.

These four discrepancy measures provided different types of information regarding the robustness of the three methods under the various missing data assumptions. We expected some of the methods to be more sensitive to the assumptions than the others.

In addition, we investigated how the filled-in data differed across the sets of missing data assumptions. This included an examination of the distributions of the filled-in missing data values, as summarized by their mean, standard deviation, skewness coefficient, kurtosis coefficient, and the 2.5th, 50th and 97.5th percentiles, as well as their correlations with the anchor test and the other tests across *P* or *Q*.

Note that all our data examples involve external anchors. However, our methodology also applies to internal anchors, which will be investigated in future.

5. Example 1: The First Pseudo-Test

The original data for this example came from one form of a licensing test for prospective teachers. The test form included 119 multiple-choice items, about equally divided among four content areas—language arts, mathematics, social studies, and science. The original form had been used at two test administrations, and the two examinee populations played the role of *P* and *Q* in our analysis.

The item responses from the original test were used to construct two pseudo-tests, *X* and *Y*. A pseudo-test consisted of a subset of the test items from the original 119-item test, and the score on the pseudo-test for an examinee was found from the responses of that examinee to the items in the pseudo-test. The pseudo-tests *X* and *Y* each contained 44 items, 11 items from each of the four content areas. Tests *X* and *Y* had no items in common and were made parallel in content. A set of 24 items (6 from each content area) were selected to be representative of the original test and to serve as the external anchor, *A*. This anchor had no items in common with either *X* or *Y*. The mean percent correct on the anchor test approximately equaled that for the 119-item original test. Further details on the construction of these pseudo-tests can be found in Holland et al. (2008).

Table 4

Statistics for the Scores on the Total and Pseudo-Tests for* P, Q, *and* P + Q *for the First Pseudo-Test Example

Test population	*N*	Original 119-items test Mean (*SD*s)	APC	*X* 44 items Mean (*SD*s)	APC	*Y* 44 items Mean (*SD*s)	APC	*A* 24 items Mean (*SD*s)	APC
P	5,168	82.3 (16.0)	.69	35.1 (5.7)	.80	26.6 (6.7)	.60	16.0 (4.2)	.67
Q	4,237	86.2 (14.2)	.72	36.4 (4.8)	.83	28.0 (6.3)	.64	17.0 (3.9)	.71
P + *Q*	0,405	83.9 (15.4)	.70	35.6 (5.4)	.81	27.2 (6.6)	.62	16.4 (4.1)	.68

Note. APC = average proportions correct.

Table 4 gives the sample sizes, means, *SD*s, and average proportion correct for the scores on *X*, *Y*, and *A* for the examinees in *P*, *Q,* and the combined group. *X* was constructed to be considerably easier than *Y*. For example, on *Q*, the mean score for *X* is larger than the mean score for *Y* by 133% of the *SD* of *Y*. In addition, *Q* is more able than *P* with a mean *A* score that is approximately a quarter of a *P* + *Q*-*SD* higher than *P*. The correlation between *X* and *A* is 0.78 and 0.74 respectively in *P* and *Q*. The correlation between *Y* and *A* is 0.79 and 0.76 respectively in *P* and *Q*.

These pseudo-tests were designed to produce an equating problem for which solutions would be nonlinear and for which the different equating methods would be expected to give different results. The large difference in difficulty between *X* and *Y* ensured that the equating functions would be nonlinear and the relatively large difference in the test performance of *P* and *Q* was supposed to ensure that CE and PSE would produce different results. These choices were made to provide a sharp comparison between the equating methods.

Because all the examinees in *P* and *Q* took all the 119 items on the original test, all of the examinees in *P* and *Q* has scores for *X*, *Y*, and *A*. In order to mimic the structure of the NEAT

design, we pretended that scores on X were not available for the examinees in Q and that scores on Y were not available for the examinees in P.

We computed the true equating functions for these data under the CE, PSE, and IRT OSE missing data assumptions as described in Section 3. In addition, because the data that are usually missing in the NEAT design are, in fact, available for the pseudo-test data, we computed a fourth true equating function using the real data. One way of viewing the pseudo-test data for X in Q and Y in P is as a fourth way of filling in the missing data, using real data rather than one of the missing data assumptions.

Figure 1 compares the estimated equating functions with the true equating functions when the missing data marginal approach was used to compute the true equating function. In figures like Figure 1, we truncate the range on the horizontal axis to lie between the 1st and 99th percentiles of the distribution of X.

Each panel in Figure 1 shows the bias, that is, the differences, between the estimated equating functions from the CE, PSE, and IRT OSE methods and one of the true equating functions (mentioned in the title of the panel). For example, the top left panel shows the bias for the CE, PSE, and IRT OSE methods when the true equating function is obtained under the CE assumptions. Note that the bias curve for the CE method in this panel is the horizontal zero-line because the bias for CE with the CE true equating function is 0 at all score points by Theorem 1. Figure 1 shows that each equating method has essentially no bias when the true equating function is computed using the missing data assumptions of that method.

The results were very similar when the filled-in missing data values were used instead of the missing data marginals to compute the true equating functions—the results are shown in Figure 2.

The dotted vertical lines in the four panels of Figures 1 and 2 mark five percentiles (5th, 25th, 50th, 75th, and 95th) of the distribution of the X scores in the full population $P + Q$. Figures 1 and 2 show that, in this example, the bias for each method is never very large, ranging only up to one-half of a raw score point above the 25th percentile. It also shows that the CE and IRT OSE methods are quite similar to each other, especially between the 5th and 95th quantiles. The fourth panels of Figures 1 and 2 show how close each estimated equating function is to the true equating function found using the real data.

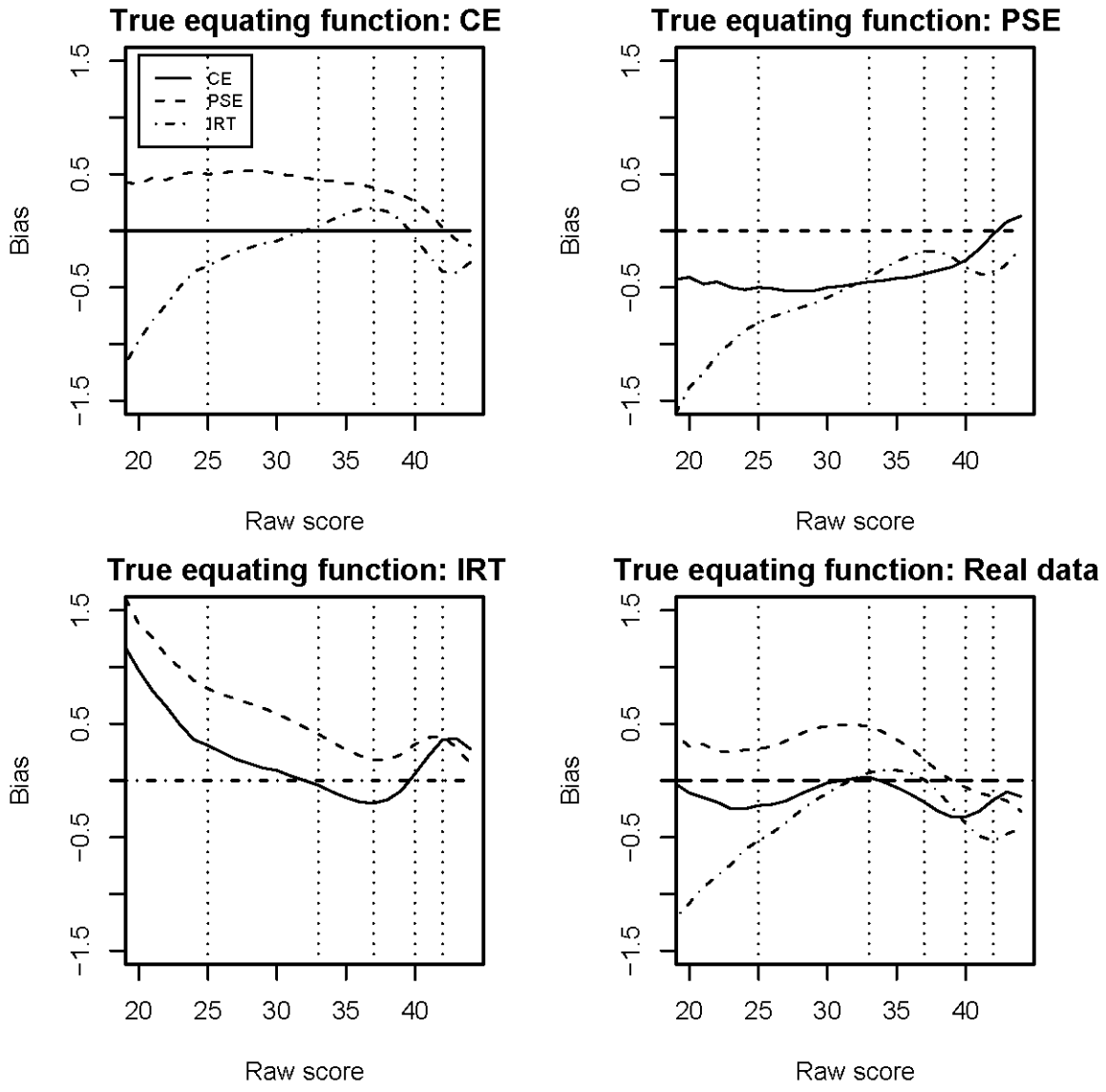

Figure 1. **Bias (differences between the estimated equating functions and the true equating function) for the first pseudo-test example when the missing data marginals are used to compute the true equating function.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

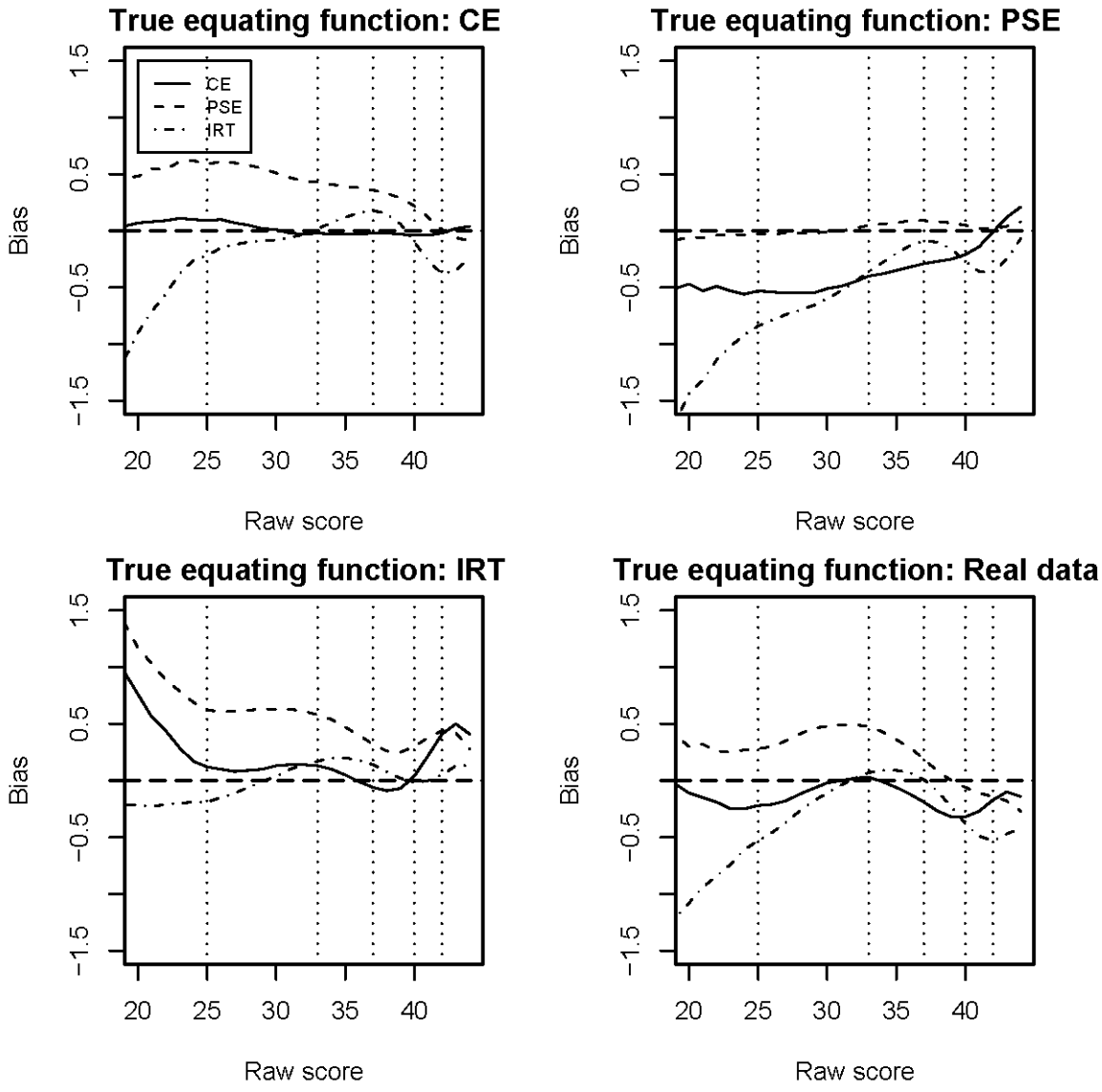

Figure 2. **Bias (differences between the estimated equating functions and the true equating function) for the first pseudo-test example when simulating the missing data values method is used to compute the true equating function. The horizontal line of long dashes denotes a difference of zero.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

Table 5 contains the four discrepancy measures, described in Section 4, comparing the estimated and true equating functions for each cell illustrated in Table 3. Note that entries are effectively zero when the true and estimated equating functions are computed using the same method.

Table 5

Summary of the Differences Between the Estimated and True Equating Functions for the Three Equating Methods for the First Pseudo-Test Example

Estimated equating function	RMSD	Assumptions behind the true equating function				Sum
		CE	PSE	IRT OSE	Real data	
CE	Entire range	0	0.40	0.31	0.20	0.91
	Lower 25%	0	0.26	0.28	0.08	0.62
	Middle 50%	0	0.30	0.11	0.17	0.58
	Upper 25%	0	0.04	0.09	0.07	0.20
PSE	Entire range	0.40	0	0.52	0.32	1.24
	Lower 25%	0.26	0	0.46	0.23	0.95
	Middle 50%	0.30	0	0.21	0.21	0.72
	Upper 25%	0.04	0	0.12	0.04	0.20
IRT OSE	Entire range	0.31	0.52	0	0.35	1.18
	Lower 25%	0.28	0.46	0	0.29	1.03
	Middle 50%	0.11	0.21	0	0.13	0.45
	Upper 25%	0.09	0.12	0	0.16	0.37

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating, RMSD = root mean squared difference.

To focus on the question of the robustness of the three equating methods across the several choices of true equating function, Table 5 shows the sum of the RMSD values across the four columns for each method. The entries in each cell of Table 5 represent RMSD values for different ranges of the X score distribution and give different views of the robustness of each equating method. For the entire score range, the ordered (low to high) RMSD sums are CE = 0.91, IRT OSE = 1.18, and PSE = 1.24. For the lower 25% of the score range, the ordered sums are CE = 0.62, PSE = 0.95, and IRT OSE = 1.03. For the middle 50% of the score range, the ordered sums are

IRT OSE = 0.45, CE = 0.58, and PSE = 0.72. For the upper 25% of the score range, the ordered sums are CE = 0.20, PSE = 0.20, and IRT OSE = 0.37.

Table 5 suggests that the CE method is the most robust among the three methods with respect to overall RMSD and the RMSD restricted to the lowest scoring and highest scoring examinees. The IRT OSE method is the most robust for the middle 50% of the score range and is the least robust for the lowest and highest scoring examinees. For this data set, the PSE method appears to be the least robust of the three methods; at best PSE ties with CE only for the highest scoring examinees.

Next we will compare the filled-in missing data for the three sets of missing data assumptions to the scores that are actually observed for *X* in *Q* and for *Y* in *P*. Table 6 displays a variety of summary statistics to facilitate this comparison.

Table 6

Summary Statistics for the Real Data and the Filled-In Missing Data for X in Q and Y in P for the First Pseudo-Test Example

Source of filled-in data	Mean (*SD*)	Skew	Kurt	2.5th % ile	Median	97.5th % ile	Correl with *A*	Correl of *X* and *Y*
				Y in *P*				
Real data	26.6 (6.7)	-0.10	-0.55	14	27	39	0.79	0.79
CE	26.4 (6.7)	-0.12	-0.51	13	26	39	0.79	0.61
PSE	26.8 (6.5)	-0.16	-0.52	14	27	38	0.77	0.60
IRT OSE	26.4 (6.7)	-0.24	-0.45	13	27	38	0.79	0.80
				X in *Q*				
Real data	36.4 (4.8)	-1.09	1.54	25	37	43	0.74	0.76
CE	36.4 (5.0)	-1.16	1.54	24	38	43	0.75	0.57
PSE	36.2 (5.1)	-1.08	1.18	23	37	43	0.76	0.57
IRT OSE	36.5 (4.9)	-1.04	1.15	24	37	43	0.75	0.77

Note. Skew = skewness coefficient, Kurt = kurtosis coefficient, Correl = correlation, % ile = percentile, CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

As shown in Table 6, all three missing data assumptions produce distributions of filled-in missing data values that are very similar to each other and to the real data. This similarity extends to means, *SD*s, skewness, kurtosis, the 2.5th, 50th, and 97.5th percentiles and to the correlation of the filled in values with *A*. The mean and *SD* of the CE and IRT OSE method are slightly closer, compared to the PSE method, to those for the real data. However, the small differences between the three methods actually reflect real differences rather than random noise. To assess this, we reran the program that filled in the missing data with a different random seed. The results were practically the same as the ones reported in Table 6. This stability is due to the large samples involved and to the fact that the sampling is conditional (i.e., stratified) on *A* rather than being purely random. When we applied the paired t-test to the difference between the real data values for each examinee and the filled-in missing data values for that examinee, all of the differences in the means were statistically significant at less than 0.005 level with the exception of CE for *X* in *Q*, which had a *p*-value of 0.48.

As another check on the stability of the small differences we saw between the equating methods, we computed the standard error of equating difference (SEED) as described in von Davier et al. (2004b) for the difference between CE and PSE using the Gaussian kernel method of continuization. The difference was well outside of the 2 SEED band for most of the score range indicating that, while small, the differences between CE and PSE, at least, were not due to random sample fluctuations. Still, what is striking about Table 6 is the similarity between the missing data distributions and those of the real data.

One important feature in Table 6 is that the IRT OSE filled-in missing data values are much better at reproducing the correlation between *X* and *Y* than the filled-in values from CE or PSE (see last column of Table 6). In fact, both in *P* and in *Q*, these correlations are much smaller for PSE and CE than they are for either the real data or the IRT OSE filled-in missing data. This feature is a direct outcome of conditioning on *X* and *A* in (33) and on *Y* and *A* in (35) for IRT OSE, whereas for CE and PSE, the conditioning is only on *A*. It is interesting that even with this advantage, IRT OSE does not agree with the real data any more than the other methods. In particular, CE does as well as IRT OSE or better at reproducing the summary statistics of the real data in Table 6. The slight superiority of CE at reproducing the real data was also shown in a different analysis of these same data in Holland et al. (2008).

6. Example 2: The Second Pseudo-Test

The pseudo-tests in this example were constructed in a similar way to the first example, using data from two administrations of a different form of the same testing program involved in our first example; these data were used in Puhan, Moses, Grant, and McHale (2008). This form had 120 multiple-choice items. Two pseudo-tests, *X* and *Y*, both 48 items long, and an external anchor *A*, 24 items long, were created from the original test. As before, this anchor had no items in common with either *X* or *Y*. Table 7 summarizes the data.

This example is less extreme than the one in Section 5. While the pseudo-tests differ substantially in difficulty with *X* being harder than *Y*, the difference is not as large as in the previous section. For example, on *Q*, the mean score on *X* is smaller than the mean score on *Y* by nearly 43% of the *SD* of *Y*. Similarly, *Q* is more able than *P* with a mean *A* score that is higher than *P* by only 14% of the *SD* of *P* + *Q*. Nevertheless, these differences were expected to be large enough to make the equating functions have a significant nonlinear component and to cause CE and PSE to differ. The correlation between *X* and *A* is 0.74 and 0.73 respectively in *P* and *Q*. The correlation between *Y* and *A* is 0.72 and 0.71 respectively in *P* and *Q*.

We used the data in this example in exactly the same way as we did for the first pseudo-test example. Figure 3, which is similar to Figure 1, compares the estimated equating functions with the true equating functions obtained using the missing data marginals approach.

Table 7

Statistics *for the Scores on the Pseudo-Tests on* P, Q, *and* P + Q, *for the Second Pseudo-Test Example*

		X 48 items		*Y* 48 items		*A* 24 items	
Test population	*N*	Means (*SD*s)	AVP	Means (*SD*s)	AVP	Means (*SD*s)	AVP
P	6,469	29.4 (6.7)	.61	32.2 (6.4)	.67	15.3 (3.5)	.64
Q	6,580	30.8 (6.3)	.64	33.4 (6.0)	.70	15.8 (3.4)	.66
P + *Q*	13,049	30.1 (6.5)	.63	32.8 (6.2)	.68	15.5 (3.5)	.65

Note. AVP = average proportions correct.

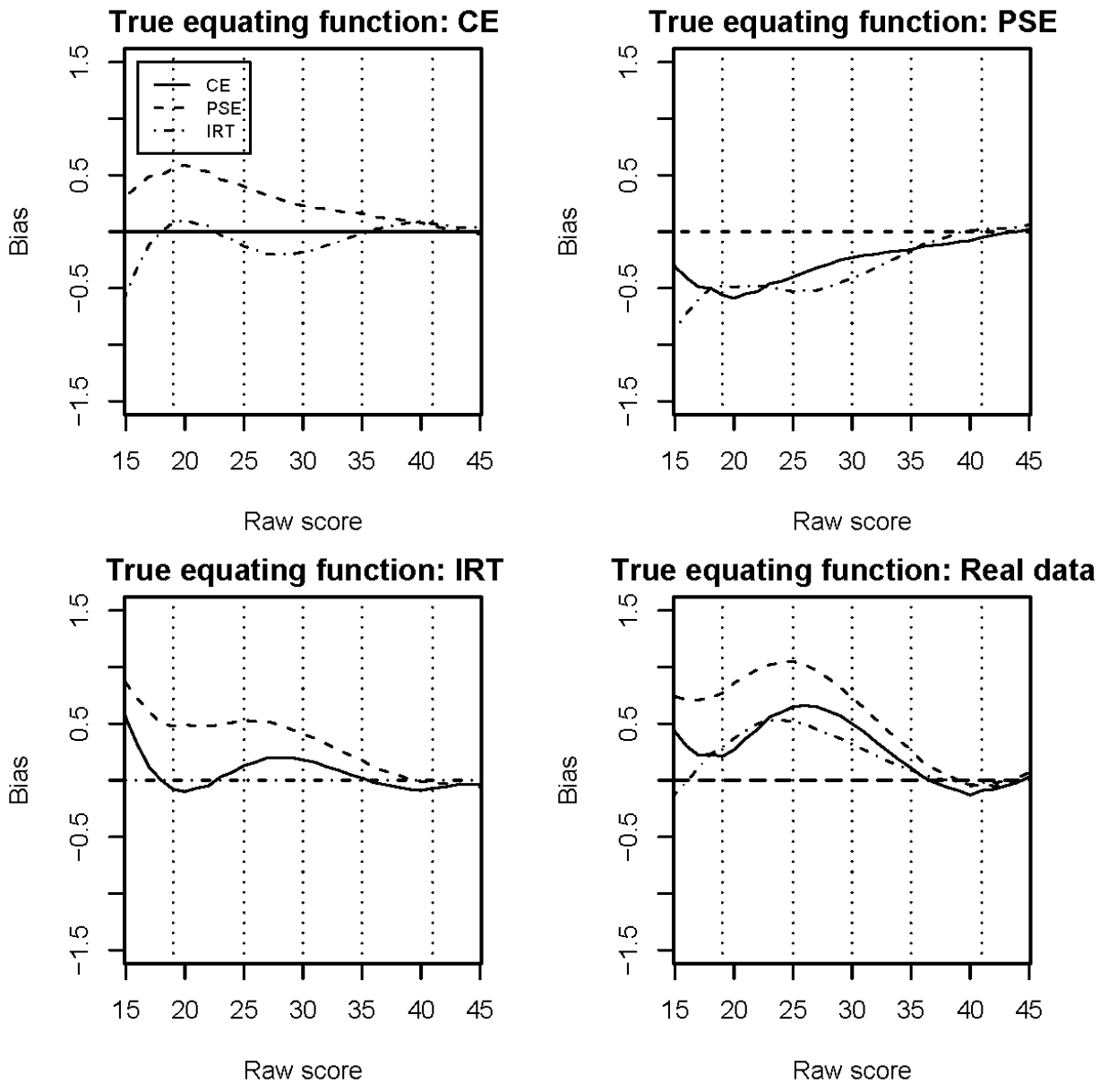

Figure 3. **Bias (differences between the estimated equating functions and the true equating function) for the second pseudo-test example when the missing data marginals are used to compute the true equating function.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

Figure 3 demonstrates that the estimated and true equating methods do not differ very much, rarely exceeding half of a raw score point above the 25th percentile. It also shows that the CE and IRT OSE methods perform similarly and are closer to the real data true equating function than is the PSE method between the 5th and 95th percentiles of the *X* score distribution.

Table 8, similar to Table 5, shows the summary measures of the discrepancies between the estimated and true equating functions.

To focus on the question of the overall robustness of the three equating methods, we again summed the RMSD values across the four columns for each method. For the entire score range, the ordered RMSD sums are IRT OSE = 0.96, CE = 0.97, and PSE = 1.47. For the lower 25% of the score range, the ordered sums are CE = 0.66, IRT OSE = 0.66, and PSE = 0.99. For the middle 50% of the score range, the ordered sums are IRT OSE = 0.68, CE = 0.68, and PSE = 1.08. For the upper 25% of the score range, the ordered sums are IRT OSE = 0.06, CE = 0.09, and PSE = 0.11.

Table 8

Summary of the Differences Between the Estimated and True Equating Functions for the Three Equating Methods for the Second Pseudo-Test Example

Estimated equating function	RMSD	The assumptions behind the true equating function				Sum
		CE	PSE	IRT OSE	Real data	
CE	Entire range	0	0.32	0.20	0.45	0.97
	Lower 25%	0	0.25	0.16	0.25	0.66
	Middle 50%	0	0.20	0.11	0.37	0.68
	Upper 25%	0	0.04	0.02	0.03	0.09
PSE	Entire range	0.32	0	0.42	0.73	1.47
	Lower 25%	0.25	0	0.28	0.46	0.99
	Middle 50%	0.20	0	0.31	0.57	1.08
	Upper 25%	0.04	0	0.03	0.04	0.11
IRT OSE	Entire range	0.20	0.42	0	0.34	0.96
	Lower 25%	0.16	0.28	0	0.22	0.66
	Middle 50%	0.11	0.31	0	0.26	0.68
	Upper 25%	0.02	0.03	0	0.01	0.06

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating, RMSD = root mean squared difference.

In Table 8, both the IRT OSE method and the CE method are the most robust with respect to the RMSD measures. The PSE is the least robust among the three methods. This result is partly due to its poor performance for the real data set. The bottom right panel in Figure 3 shows that the bias of the PSE method is more than that of the other two methods, especially below the 75th percentile of the total score range.

In Table 9, we compare the distributions of the filled-in missing data values with the real data for *X* in *Q* and for *Y* in *P*.

Table 9

Summary Statistics for the Real Data and the Filled-In Missing Data for* X *in* Q *and* Y *in* P *for the Second Pseudo-Test Example

Source of filled-in data	Mean (*SD*)	Skew	Kurt	2.5th % ile	Median	97.5th % ile	Correl with *A*	Correl of *X* and *Y*
				Y in *P*				
Real data	32.2 (6.4)	-0.29	-0.40	19	33	43	0.72	0.81
CE	32.4 (6.2)	-0.37	-0.30	19	33	43	0.73	0.54
PSE	32.8 (6.1)	-0.40	-0.24	20	33	43	0.72	0.53
IRT OSE	32.5 (6.3)	-0.40	-0.21	19	33	43	0.73	0.81
				X in *Q*				
Real data	30.8 (6.3)	-0.19	-0.44	18	31	42	0.73	0.79
CE	30.5 (6.5)	-0.10	-0.53	18	31	42	0.74	0.53
PSE	30.1 (6.6)	-0.04	-0.61	18	30	42	0.73	0.52
IRT OSE	30.5 (6.4)	-0.28	-0.36	17	31	41	0.72	0.80

Note. Skew = skewness coefficient, Kurt = kurtosis coefficient, Correl = correlation, % ile = percentile, CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

Table 9 again shows that all three missing data assumptions reproduce the distribution of the real data about equally well. Again, some small differences exist, but the means, *SD*s, skewness, kurtosis, and the three percentiles of the filled-in missing data values are remarkably close to the values for the real data values. The means and *SD*s of CE and IRT OSE are a bit closer to the real data values than those of the PSE, but the differences are small. Again, IRT OSE is better at reproducing the correlation between *X* and *Y* than the other two methods for the reasons mentioned in Section 5.

7. Example 3: The Admissions Test

The data in Table 10 are from the verbal measure of a large admissions testing program that uses the NEAT design in its operational equating process. The data were collected from two test forms, a new form *X* that we equated to an old form *Y*. Both test forms had 78 multiple-choice items divided among three item types. *X* and *Y* were designed to be parallel and similar in difficulty. Unlike the two previous examples, there were no real data for *X* in *Q* or for *Y* in *P*. A 35-item, multiple-choice, external anchor test, *A,* was also included. Table 10 summarizes the data.

Unlike the two previous examples, no direct way tells which of these test forms was harder and by how much. This information emerged from the results of the test equatings. We see in Table 10 that *X* is slightly harder than *Y*. Population *P* is less able than *Q*. The average *A* score in *P* is one fourth of the *SD* of *P* + *Q* below that in *Q*. The correlation between *X* and *A* in *P* is 0.90, and the correlation between *Y* and *A* in *Q* is 0.91.

Table 10

Statistics for the Scores on the New and Old Tests and the Anchor Test on* P *and* Q *for the Admissions Test Example

		X 78 items		*Y* 78 items		*A* 35 items	
Test population	*N*	Mean (*SD*)	AVP	Mean (*SD*)	AVP	Mean (*SD*)	AVP
P	5,226	41.4 (15.0)	.53	-	-	18.7 (7.3)	.53
Q	11,297	-		45.7 (14.7)	0.59	20.5 (7.1)	.59
P + *Q*	16,523	-		-		19.9 (7.2)	.57

Note. AVP = average proportion correct.

Figure 4 compares the estimated and true equating functions under the three sets of missing data assumptions when the missing data marginals were used to compute the true equating functions.

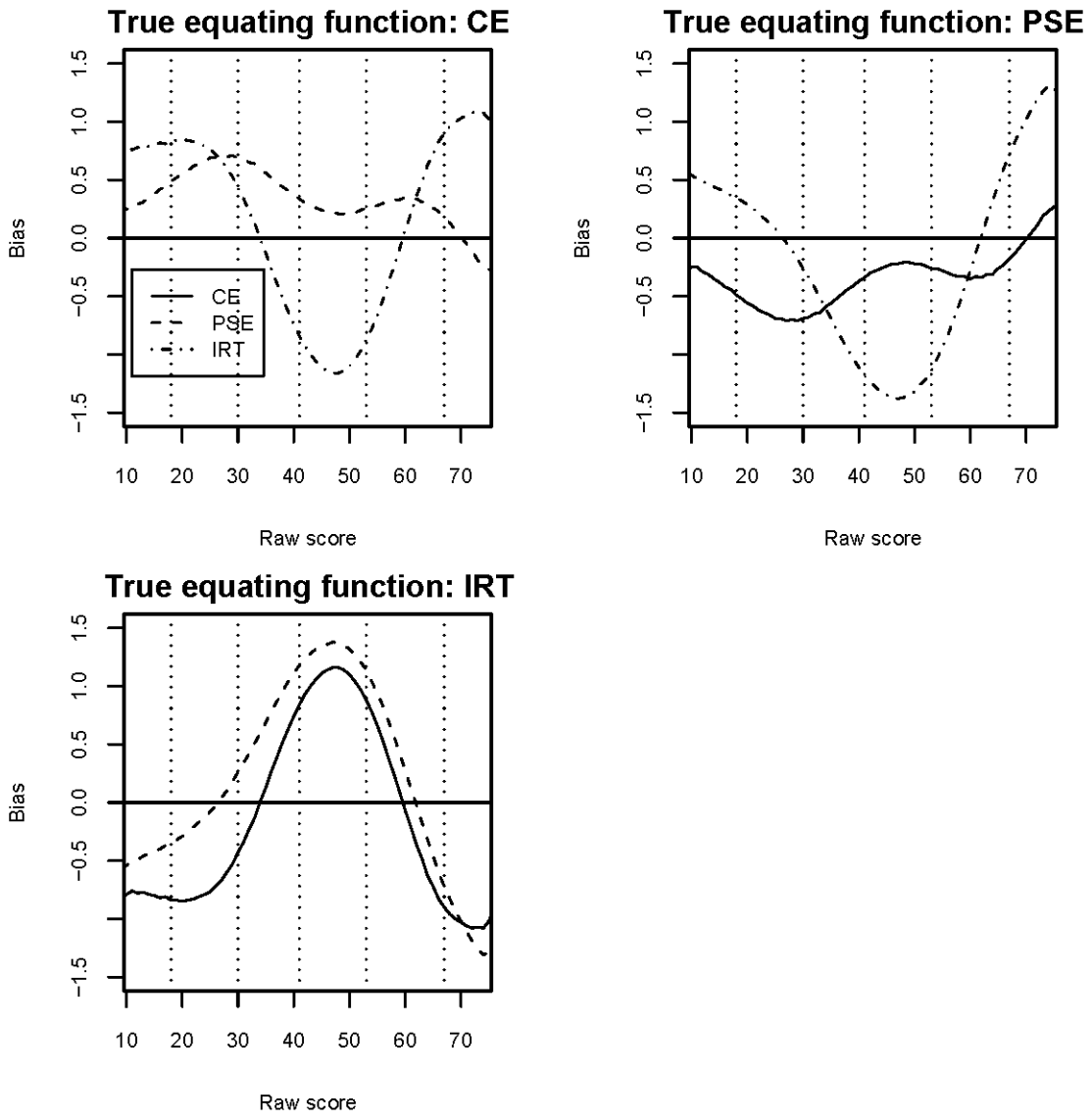

Figure 4. **Bias (differences between the estimated equating functions and the true equating function) for the admissions test example when the missing data marginals are used to compute the true equating function.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

Figure 4 shows that, in this example, equating functions for CE and PSE are more similar to each other and differ from the IRT OSE function. Moreover, in this example, the differences between IRT OSE and the other two methods exceed one raw score point in important regions of the score range. Because we have no real data for comparison, it is unclear whether this indicates that PSE and CE are wrong or that IRT OSE is wrong. Nonetheless, this worst case has produced differences that are only at the edge of concern for a real testing program.

Table 11 (which is similar to Tables 5 and 8 but which does not have a column for real data) shows the RMSD measures. For the entire score range, the ordered RMSD sums are CE = 1.22, PSE = 1.31, and IRT OSE = 1.63. For the lower 25% of the score range, the ordered sums are PSE = 0.45, IRT OSE = 0.54, and CE = 0.69. For the middle 50% of the score range, the ordered sums are CE = 0.90, PSE = 1.08, and IRT OSE = 1.36. For the upper 25% of the score range, the ordered sums are CE = 0.43, PSE = 0.47, and IRT OSE = 0.64.

Table 11

Summary of the Differences Between the Estimated and True Equating Functions for the Three Equating Methods for the Admissions Test Example

The estimated equating function	RMSD	The assumptions behind the true equating function			Sum
		CE	PSE	IRT OSE	
CE	Entire range	0	0.45	0.77	1.22
	Lower 25%	0	0.30	0.39	0.69
	Middle 50%	0	0.31	0.59	0.90
	Upper 25%	0	0.13	0.30	0.43
PSE	Entire range	0.45	0	0.86	0.31
	Lower 25%	0.30	0	0.15	0.45
	Middle 50%	0.31	0	0.77	1.08
	Upper 25%	0.13	0	0.34	0.47
IRT OSE	Entire range	0.77	0.86	0	1.63
	Lower 25%	0.39	0.15	0	0.54
	Middle 50%	0.59	0.77	0	1.36
	Upper 25%	0.30	0.34	0	0.64

Note. The four values in each cell denote the RMSDs for the entire range, for lower 25%, for middle 50%, and for upper 25%. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating, RMSD = root mean squared difference.

In Table 11, in terms of the RMSD values, CE and PSE show the most robustness but differ as to where along the score range they show this. The CE method is the most robust with respect to the overall RMSD, middle 50% of the examinees, and the highest scoring 25% of the examinees; the PSE method is the least robust with respect to the lowest scoring 25% of the examinees. The IRT method shows the least robustness in this example.

Table 12, which is similar to Tables 6 and 9 but does not have a column for the real data, displays the summary measures of the distributions of the filled-in missing data values for the three sets of missing data assumptions.

Table 12

Summary Statistics for the Real Data and the Filled-In Missing Data for* X *in* Q *and* Y *in* P *for the Admission Test Example

Source of the filled-in data	Mean (*SD*)	Skew	Kurt	2.5th % ile	Median	97.5th % ile	Correl with *A*	Correl of *X* and *Y*
				Y in *P*				
CE	42.0 (15.0)	0.06	-0.79	15	42	70	0.91	0.82
PSE	42.3 (15.0)	0.04	-0.81	15	42	70	0.91	0.81
IRT OSE	41.8 (15.0)	0.10	-0.75	15	41	70	0.91	0.94
				X in *Q*				
CE	45.2 (14.7)	-0.06	-0.74	18	45	71	0.90	0.82
PSE	44.7 (14.8)	-0.07	-0.71	16	45	71	0.90	0.81
IRT OSE	45.3 (14.6)	-0.20	-0.71	17	46	70	0.91	0.94

Note. Skew = skewness coefficient, Kurt = kurtosis coefficient, Correl = correlation, % ile = percentile, CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

8. Discussion, Conclusions, and Recommendations

Discussion

The similarities and differences among the three examples are worth summarizing.

First, in all three examples, the tests to be equated, *X* and *Y*, are parallel in content and numbers of items. The anchor tests are also similar in content to the tests to be equated and have difficulty levels that are between those of *X* and *Y*. In these examples, the test-anchor-test correlations are all above 0.70, sometimes quite substantially. These examples are, therefore, all cases in which test equating is likely to be successful.

Second, in all three cases, large *A* score differences appear between *P* and *Q* (the difference in means lies between 14% to 25% of an *SD*). This condition usually ensures that CE and PSE will produce different results.

Third, the difference in test difficulty between *X* and *Y* varies considerably over the three examples (from 133% to 43% to 7% of a *Q*-*SD* difference). A large difficulty difference usually ensures the need for a nonlinear equating function; a linear equating function might be acceptable when the test differences in difficulty are small.

To show the effect of test difficulty differences on equating functions, Figures 5 to 7 plot the estimated equating functions for the three examples. The functions in Figures 5 and 6 exhibit a definite curvilinearity, while Figure 7 shows that the relationships are nearly linear. In Figures 5 and 6, we plotted the 45-degree line that would be the equating function if the two tests were equally difficult. The 45-degree line was in the middle of the graphs of the equating functions in Figure 7 and was omitted.

Fourth, the three examples have different test and anchor-test lengths (44 and 24, 48 and 24, and 78 and 35, respectively).

Fifth, the three examples also differ with respect to the relative samples sizes from *P* and *Q*. This difference means that the choice of *w* that applies when *P* and *Q* are pooled to form *P* + *Q* differs in the three cases. In Example 1, $w = 6168/10405 = 0.59$. In Example 2, $w = 6469/13049 = 0.50$. In Example 3, $w = 5226/16523 = 0.32$. Previous research suggests that the choice of *w* has no influence on the results of CE and only a minor influence on the results of PSE. Hence, *w* probably has little influence on IRT OSE, but this supposition has yet to be established. In our opinion, the fact that *w* varies in these examples has little consequence for the differences that we see in the results.

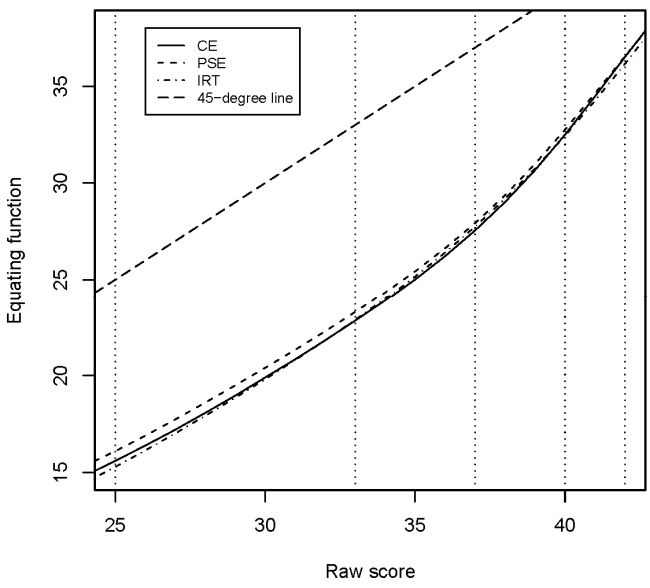

Figure 5. **The three estimated equating functions for the first pseudo-test example.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

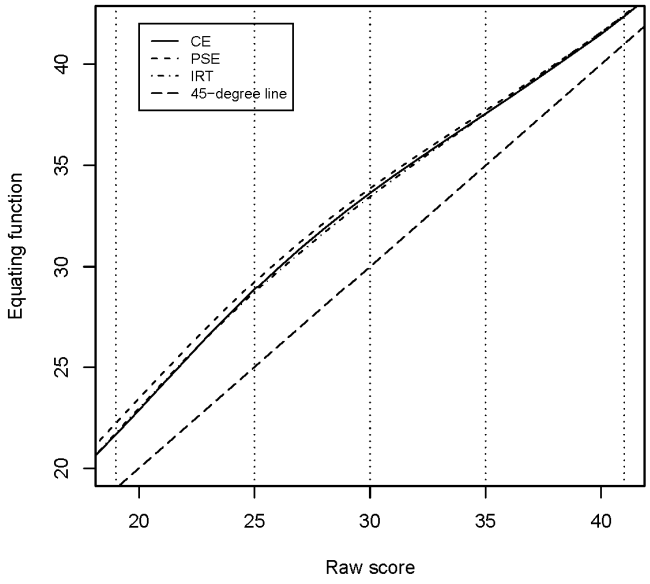

Figure 6. **The three estimated equating functions for the second pseudo-test example.**

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

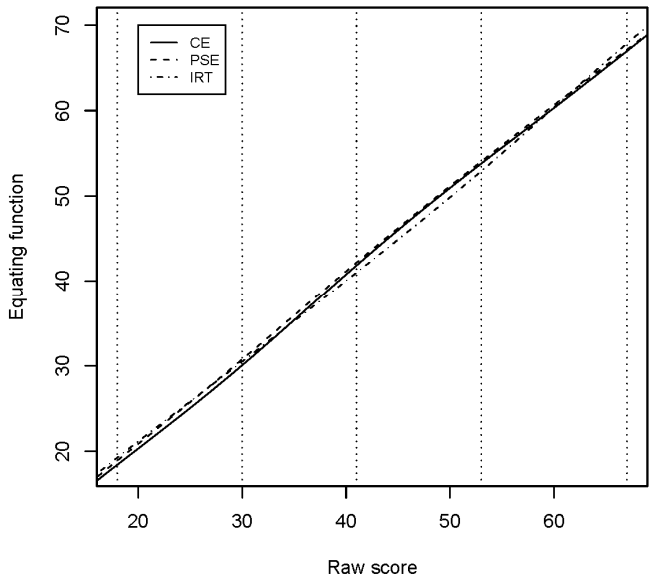

Figure 7. The three estimated equating functions for the admissions test example.

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

Thus, the three examples cover a range of situations that (a) are all plausible circumstances where test equating is likely to be successful and (b) are expected to vary from linear to nonlinear equating functions. Moreover, they ought to show differences among the three equating methods. An examination of the three tables of bias and RMSD values, Tables 5, 8, and 11, do not show one of the methods consistently dominating the other two in terms of their robustness against the various missing data assumptions we used. To get a simple summary of how each equating method fared across the three examples, we examined the three missing data assumption columns of Tables 5, 8, and 11 (CE, PSE, and IRT OSE) separately from the real data columns in Tables 5 and 8. To simplify the numbers for each method, we ranked its bias or RMSD values compared to the other two methods separately for each missing data assumption. For its own missing data assumption, a method got a rank of 0 (because the estimated equating function and the true equating function are the same for any method). For any other assumption, a method got a rank of 1 or 2 depending on whether the value for the method was the smaller or larger of the two nonzero values. For example, in Table 5 (the first pseudo-test) under the CE assumption, the overall RMSD

is 0 for the CE method, 0.31 for the IRT OSE method, and 0.40 for the PSE method. Thus, under the CE assumption, the rank for the overall RMSD of CE is 0, the rank for the overall RMSD of IRT OSE is 1, and the rank for the overall RMSD of PSE is 2. The complete set of these ranks is displayed in the three panels of Table 13 (the zero ranks are omitted).

Table 13

The Ranks of the Root Mean Squared Difference (RMSD) Measures in Tables 3, 6, and 9 Organized by Equating Method, Example, and Missing Data Assumptions

	First example			Second example			Third example		
Assumptions	CE	PSE	IRT	CE	PSE	IRT	CE	PSE	IRT
CE method									
Overall RMSD		1	1		1	1		1	1
RMSD for < 25% ile		1	1		1	1		2	2
RMSD for 25-75% ile		2	1		1	1		1	1
RMSD for > 75% ile		1	1		2	1		1	1
PSE method									
Overall RMSD	2		2	2		2	1		2
RMSD for < 25% ile	1		2	2		2	1		1
RMSD for 25-75% ile	2		2	2		2	1		2
RMSD for > 75% ile	1		2	2		2	1		2
IRT OSE method									
Overall RMSD	1	2		1	2		2	2	
RMSD for < 25% ile	2	2		1	2		2	1	
RMSD for 25-75% ile	1	1		1	2		2	2	
RMSD for > 75% ile	2	2		1	1		2	2	

Note. % ile = percentile, CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

In Table 13, the more times 1 appears in a panel, the greater the robustness of the corresponding method (i.e., the smaller it deviates from the true equating function for the sets of missing data assumptions that do not favor it). Each panel in Table 13 has 24 nonzero entries. For CE there are twenty 1s and four 2s. For PSE there are eight 1s and sixteen 2s. For IRT OSE there

are nine 1s and fifteen 2s. From this perspective, CE appears to exhibit more robustness than the other two methods across the three sets of missing data assumptions, and PSE exhibits less robustness than the other two methods; IRT OSE performs in between.

For the two pseudo-test examples, a table similar to Table 13 can be made for the RMSD values under the real data columns of Tables 5 and 8. Table 14 gives the rankings (now 1, 2, or 3) of the three values for CE, PSE, and IRT OSE for each measure and the two examples.

Table 14

The Ranks of the Root Mean Squared Difference (RMSD) Measures for the Real Data Columns of Tables 3 and 6 Organized by Equating Method and Example

Equating method	CE		PSE		IRT OSE	
	First example	Second example	First example	Second example	First example	Second example
Overall RMSD	1	2	2	3	3	1
RMSD for < 25% ile	1	2	3	3	2	1
RMSD for 25-75% ile	2	2	3	3	1	1
RMSD for > 75% ile	2	2	1	3	3	1

Note. % ile = percentile, CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

In Table 14, an equating method with smaller ranks is more able than one with higher ranks to reproduce the true equating function that corresponds to the real data. To get a quantitative comparison, we summed the ranks for each equating method in Table 14. The sums are 14 for CE, 21 for PSE, and 13 for IRT. In the second example, the IRT OSE method did the best according to all four of our measures; this explains its success in this quantitative comparison. The IRT OSE and CE methods do noticeably better than PSE in this comparison.

Putting the results of Tables 13 and 14 together, we see that the CE method appears to have an edge over the other two methods with respect to robustness to the several plausible missing data assumptions. In particular, the PSE and the IRT OSE methods do not appear to do consistently as well with respect to this criterion.

What is quite clear in these examples, however, is how similar the three methods are even in circumstances (i.e., large differences between *P* and *Q*) where we thought that they would be

substantially different. This leads us to consider a comparison of Tables 6, 9, and 12. The striking observation is how the various missing data assumptions lead to very similar distributions of the missing data, and how closely these all agree with the real data when these are available. Apparently, the very tiny differences seen across the missing data assumptions in these three tables are the sources of the final small differences in the resulting equating functions.

How do the filled-in missing data values and the real data relate to each other? From Tables 6, 9, and 12, we see that the marginal distributions are very close, but probably different enough to result in the small differences that are observed in the equating functions. What about the correlations between the filled-in missing data values and the real data? These correlations are given in Table 15.

Table 15

Correlations Between the Filled-In Missing Data Values for the Various Missing Data Assumptions and the Real Data for the Three Examples

	Y in *P*				*X* in *Q*			
	Real data	CE	PSE	IRT OSE	Real data	CE	PSE	IRT OSE
First pseudo-test example								
Real data	-	0.61	0.62	0.72	-	0.57	0.56	0.68
CE	-	-	0.61	0.62	-	-	0.59	0.57
PSE	-	-	-	0.62	-	-	-	0.58
Second pseudo-test example								
Real data	-	0.53	0.52	0.69	-	0.54	0.52	0.68
CE	-	-	0.53	0.54	-	-	0.54	0.54
PSE	-	-	-	0.52	-	-	-	0.53
Admissions test example								
CE	-	-	0.53	0.54	-	-	0.54	0.54
PSE	-	-	-	0.52	-	-	-	0.53

Note. CE = chain equipercentile equating, IRT OSE = item-response-theory observed-score equating, PSE = poststratification equating.

The correlations between the filled-in missing data for CE, PSE, and IRT OSE are very similar for the second and third example, ranging from 0.52 to 0.54. The corresponding correlations for the first example are slightly higher, ranging from 0.57 to 0.62. In the first two examples, the correlations of the real data with the filled-in missing data values from IRT OSE are higher (0.68 to 0.72) than the other correlations (0.52 to 0.62) due in large measure to the extra conditioning that goes into the simulation of the IRT OSE values. It is remarkable that this higher correlation does not translate into a consistently more accurate or robust equatings using the IRT OSE method.

Conclusions and Recommendations for Future Research

In this study, we showed ways to fill in the missing data of the NEAT design so that the estimated equating function obtained by any one of the three equating methods would appear to be nearly equal to the true equating function. This illustrates an aspect of the statement in Wang et al. (2006) that "it is virtually impossible in a simulation to have a criterion that does not advantage/disadvantage one or more methods to some degree" (p. 15). Simulating data from an IRT model in an equating study can bias the results so that IRT OSE will tend to agree with an IRT-based true equating function. Our results suggest that CE will tend to be the second best in an IRT-based simulation. Examination of the IRT OSE columns of Tables 5, 8, and 11 shows that PSE has a smaller discrepancy measure in only one case, the RMSD value for the lower 25% of examinees in the third example (Table 11). However, our results show that there are many ways to fill in the missing data of the NEAT design beyond those of an IRT model, and we believe that they should all become part of the tool kit for equating studies. Hence, this paper promises to have a significant impact on future studies comparing different equating methods.

We found that the two choices of defining the true equating function, the missing data marginals approach and the filling-in the missing data values approach, give nearly the same results, at least in the types of examples we considered.

Another finding is the respectable performance of the IRT OSE method. It performs better than the two other methods when the true equating function is obtained under the real data assumption in the second example and does not perform too much worse than the other two methods in the other cases. Thus, our research partially responds to the appeal in Kolen and Brennan (2004) to perform more research studies involving the IRT OSE method and shows that

the method is indeed worthy of further attention by practitioners, especially in testing programs that use an IRT model to report scores.

The results show a relatively close agreement between the three observed score equating methods. While several factors varied in this study, the three methods agreed much more than they disagreed. The small exception to this is the somewhat larger difference showed by IRT OSE from the other two methods for the third example.

To understand why the methods agree so well, we believe that one important factor across the examples cannot be overemphasized: In all examples, test equating was expected to be successful. The tests we equated were very similar in content and identical in test length (and therefore similar in reliability); the anchor tests were sufficiently reliable (similar to *X* and *Y* and highly correlated with them). The differences in test difficulty across the three examples, while leading to nonlinear equating functions, did not seem to matter very much as far as the equating methods we studied were concerned. The only factor in these examples that usually leads to inaccurate equating was that *P* and *Q* are definitely different in every case.

As indicated above, we believe that the three equating methods agree so well—and what differences they have are so small across the three examples—because it makes sense to equate *X* and *Y* in these examples. While it is generally true that we should expect the largest differences between CE and PSE (and possibly IRT OSE) when *P* and *Q* differ in ability, this expectation does not mean that these differences will be so large as to make the final choice of an equating method problematic. Generally, in cases where it makes sense to equate *X* and *Y*, they will agree well enough for practical purposes. Had *X* and *Y* differed substantially in content or in reliability or had otherwise violated the *Five Requirements of Test Equating* as discussed in Holland and Dorans (2006, pp. 215–216),[4] then we probably would have seen larger differences among the three equating methods. A suggestion for future research in this area then is to create pseudo-tests or find real test data examples where equating or test linking can be performed between tests that violate the assumptions of similar content and reliability or when the anchor tests are less reliable or less similar to the tests to be equated. Our research here indicates that without such departures from ideal equating circumstances, little chance exists of seeing important differences between the types of observed score equating methods studied here.

What about the robustness of the three methods to the missing data assumptions that are made? Here we are fairly confident that the CE method is a little bit more satisfactory than the

other two methods. This finding is just another in a long line of findings that show that CE methods for the NEAT design are sound and of practical value even though they appear to lack the more elegant theoretical support that characterizes PSE and IRT OSE methods. Finally, we should restate our surprise at finding that, while the IRT OSE methods of filling in the missing data of the NEAT design are the most strongly correlated with the real data when that is available, this correlation does not appear to give IRT OSE any particular advantage over the other two methods in the three examples of this study. Perhaps it will pay off when the equating situation is less satisfactory than those in our three examples. This topic is one for future research.

In view of our research, what advice can we give to those who need to equate tests using the NEAT design? First of all, this work shows, using a new criterion—robustness against variations in missing data assumptions—that the simple CE methods hold up well. While the CE method has been regarded as having "shortcomings" (Kolen & Brennan, 2004, p. 146), a large body of evidence now shows that, whatever its shortcomings, it works well in a variety of circumstances. We would encourage practitioners to consider using CE methods for the NEAT design and not to shy away from them because they appear to be too simple to be right.

More generally, our research shows that regardless of which of the three equating methods is used by practitioners for the NEAT design, the methods will tend to agree with each other when the proper conditions for equating are in place (the proper conditions for equating with anchor tests are summarized in, for example, Holland & Dorans, 2006, pp. 215–216). This agreement will hold even when the tests differ quite widely in difficulty or the two groups differ markedly on the anchor test. Looking at the agreement we found, large discrepancies among the results of CE, PSE, and IRT OSE, is possibly a sign that a problem needs further examination. In other words, it is a good idea to calculate more that one type of equating function for important tests in order to uncover problems that might not reveal themselves otherwise.

In *Uncommon Measures*, Feuer, Holland, Green, Bertenthal, and Hemphill (1999) explained to the U.S. Congress why attempts to link scores on tests that differ in content and reliability (and possibly in other things as well) will generally result in unsatisfactory linkages that do not stand up across different testing populations and across time. Our research shows that when attempts to link scores on tests are carried out under favorable circumstances, they will be successful in the sense that different equating methods will give similar results even when the tests are very different in difficulty and when the populations arising in the NEAT design differ by up to

one-quarter of an *SD* on the anchor test. Feuer et al. (1999) were hampered by the lack of extensive research and examples to show in detail what problems would arise when test equating was attempted in unsatisfactory situations. Their findings depended to a great extent on professional judgment rather than on a large array of scientific equating studies. We believe that research that extends the present set of examples to less satisfactory test equating situations would bolster the conclusions of Feuer et al. and clarify the degree to which test linking becomes problematic under less than ideal test-linking circumstances.

References

Bishop, Y. M. M., Fienberg, E. F., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

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. 71–135). New York: Academic Press.

Feuer, M., Holland, P., Green, B., Bertenthal, M., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington DC: National Academy Press.

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL model* (ETS Research Rep. No. RR-06-10). Princeton, NJ: ETS.

Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50,* 61–71.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*(4), 577–601.

Holland, P. W., & Dorans, N. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement*, *45*, 17–43.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25,* 133–183.

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design* (ETS Research Rep. No. RR-06-17). Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking* (3rd ed.). New York: Springer.

Liou, M., & Cheng, P. E. (1995). Equipercentile equating via data-imputation techniques. *Psychometrika, 60*(1), 119–136.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95.

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Puhan, G., Moses, T., Grant, M., & McHale, F. (2008). *An alternative data collection design for equating with very small samples* (ETS Research Rep. No. RR-08-11). Princeton, NJ: ETS.

Sinharay, S. (2008, September). *Chain equating versus post-stratification equating: An illustrative comparison.* Paper presented at Looking back: A conference to honor Paul Holland, Princeton, NJ.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*, 249–275.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41,* 15–32.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. New York: Springer Verlag.

Wang, T., Lee, W.-C., Brennan, R. J., & Kolen, M. J. (2006, April). *A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Notes

[1] There are approaches for filling in the missing data that do not reflect the missing data of these methods (Liou & Chen, 1995).

[2] We also used the kernel smoothing method (KSM; von Davier et al., 2004b), but it produced nearly identical results as the traditional linear interpolation method. So we do not report results for the KSM.

[3] Note that we computed $\mathrm{P}\{X = x \mid P\}$ and $\mathrm{P}\{Y = y \mid Q\}$ from the presmoothed observed bivariate distributions, as described earlier.

[4] The five requirements are (a) equal construct, (b) equal reliability, (c) symmetry, (d) equity, and (e) population invariance.

Appendix A
Raking

Raking is an iterative algorithm that produces a discrete bivariate distribution that has the association (as measured by the adjacent cross-product ratios) of one bivariate discrete distribution and the marginal distributions of two other target marginal distributions from different sources. In the application used here to produce $\{k_{xa\mathbf{Q}}\}$ in (23), the discrete bivariate distribution is the presmoothed joint distribution of X and A over P, $\{p_{xa}\}$; the two target marginal distributions are the discretized missing data marginal of X in Q, $\{f_{x\mathbf{Q}}\}$, from (22), and the presmoothed marginal of A in Q, $\{h_{a\mathbf{Q}}\}$. The result, $\{k_{xa\mathbf{Q}}\}$, is an estimate of the (unobserved) joint distribution of X and A in Q. In the original application of raking, the two discrete marginal distributions came from a census of a population for which the joint distribution was not available, and the joint distribution of the two variables was known from a sample whose marginals may have differed substantially from the population values. Raking was developed to combine these two sources of information in a consistent way. Raking is related to maximum likelihood estimation of hierarchical loglinear models for multiway contingency tables.

The raking algorithm consists of alternating row and column multiplications of the entries in the joint distribution in such a way that the result converges to a joint distribution that has the two target marginal distributions. For example, begin with the rows, then take each row in turn and multiply each entry in the row by the ratio of the target marginal value for that row to the current total for that row; the row totals of the adjusted data agree with the target marginal for the row variable. The column totals of the adjusted data, however, may not agree with the target marginal for the column variable. Thus, the next step, taking each column in turn, is to multiply each entry in a column by the ratio of the target marginal for the column variable to the current column total for that column. Now the column totals of the adjusted data agree with the target marginal for that variable, but the new row totals may no longer match their target marginals. The process iterates, alternating between the rows and the columns, and agreement on both rows and columns is usually achieved after a few iterations. The result is a joint distribution that has the target marginals for both variables. In addition, because the steps in the iterations involve only row and column multiplications, the cross-product ratios for the adjacent two-by-two tables within the larger joint distribution will be unchanged from the original joint distribution. In this sense, the final result is a joint distribution that preserves the cross-product association or

dependence structure of the original joint distribution, but it now has the desired target marginal distributions.

In our example, to obtain an estimate of the (unobserved) joint distribution of X and A in Q, the raking algorithm will iterate through the following steps (assume that the lowest and highest possible X scores are l_X and h_X, respectively, and l_A and h_A for the A scores):

1. For $x \in (l_X, h_X)$, compute, for $a \in (l_A, h_A)$, $p^*_{xa} = p_{xa} \frac{f_{xQ}}{\sum_a p_{xa}}$.

2. For $a \in (l_A, h_A)$, compute, for $x \in (l_X, h_X)$, $p^{**}_{xa} = p^*_{xa} \frac{h_{aQ}}{\sum_x p^*_{xa}}$.

3. Reset $p_{xa} = p^{**}_{xa}$ and continue until convergence.

See Bishop et al. (1975) for further discussion of raking, which is called *iterative proportional fitting* there.

Appendix B

Proof of Theorem 1

By definition, $F_T(x) = wF_P(x) + (1 - w)F_Q(x)$, so substitute $F_Q(x) = H_Q(H_P^{-1}(F_P(x)))$ from (16) to obtain

$$\begin{aligned} F_T(x) &= wF_P(x) + (1 - w)H_Q(H_P^{-1}(F_P(x))) \\ &= wH_P(H_P^{-1}(F_P(x))) + (1 - w)H_Q(H_P^{-1}(F_P(x))) \\ &= (wH_P + (1 - w)H_Q)(H_P^{-1}(F_P(x))). \end{aligned} \tag{B1}$$

Equation (B1) follows from the distributive property of function composition. Furthermore, by definition, $H_T(x) = wH_P(x) + (1 - w)H_Q(x)$, so that combining the above we have

$$F_T(x) = H_T(H_P^{-1}(F_P(x))). \tag{B2}$$

In a similar way we may show that

$$G_T(y) = H_T(H_Q^{-1}(G_Q(y))) \tag{B2}$$

or

$$G_T^{-1}(p) = G_Q^{-1}(H_Q(H_T^{-1}(p))) \tag{B3}$$

so that

$$G_T^{-1}(F_T(x)) = G_Q^{-1}(H_Q(H_T^{-1}(H_T(H_P^{-1}(F_P(x)))))) = G_Q^{-1}(H_Q(H_P^{-1}(F_P(x)))),$$

as was to be proved.