Research Report

Testing the Untestable Assumptions of the Chain and Poststratification Equating Methods for the NEAT Design

Paul W. Holland

Alina A. von Davier

Sandip Sinharay

Ning Han

Research & Development

June 2006
RR-06-17

Testing the Untestable Assumptions of the Chain and Poststratification Equating Methods for the NEAT Design

Paul W. Holland, Alina A. von Davier, Sandip Sinharay, and Ning Han
ETS, Princeton, NJ

June 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

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

Abstract

This paper focuses on the Non-Equivalent Groups with Anchor Test (NEAT) design for test equating and on two classes of observed–score equating (OSE) methods—chain equating (CE) and poststratification equating (PSE). These two classes of methods reflect two distinctly different ways of using the information provided by the anchor test for computing OSE functions. Each of the two classes includes linear and nonlinear equating methods. In practical situations, it is known that the PSE and CE methods tend to give different results when the two groups of examinees differ in ability. However, given that both methods are justified by making untestable assumptions, it is difficult to conclude which, if either, of the two equating approaches is more correct. This study compares predictions from both the PSE and the CE assumptions that can be tested in a comparable way with the data from a special study. Results indicate that both CE and PSE make very similar predictions but that those of CE are slightly more accurate than those of PSE.

Key words: Test equating, Non-Equivalent Groups with Anchor Test (NEAT) design, observed-score equating, chain equating, poststratification equating, missing data, pseudo-tests, continuization, discretization

Acknowledgments

We would like to thank our colleagues Tim Moses and Michael Walker for their careful reading of an earlier draft of this report and Tim for the additional analyses he did to clarify our work.

Introduction

Test equating methods are widely used to produce scores that are comparable across different forms of the same test, both within a year and across years. This paper focuses on the Non-Equivalent Groups with Anchor Test (NEAT) equating design and on two classes of observed–score equating (OSE) methods—chain equating (CE) and poststratification equating (PSE). PSE and CE reflect two distinctly different ways of using the information provided by the anchor test—poststratifying on the anchor to estimate the score distributions for the tests to be equated (PSE) or using the anchor test as the middle link in a chain of linking relationships (CE). Each of the two classes of methods includes both linear and nonlinear equating functions. The PSE methods include the Tucker and Braun-Holland linear methods and the nonlinear frequency estimation method (see Braun & Holland, 1982; Kolen & Brennan, 2004; Livingston, 2004). The CE methods include both the chained linear and the chained equipercentile methods. The nonlinear methods of CE and PSE have parallel versions formed by continuizing the discrete distributions by either linear interpolation (Kolen & Brennan) or by Gaussian kernel smoothing (von Davier, Holland, & Thayer, 2004a).

von Davier et al. (2004a) examined the relationship between the CE and PSE methods in the NEAT design and both were shown to be examples of OSE methods under different sets of *population invariance assumptions*. These assumptions are *untestable* using the data usually available in the NEAT design. von Davier, Holland, and Thayer (2004b) showed that under certain (idealized) conditions both CE and PSE can produce the same equating function.

In practical situations, the PSE and CE methods tend to give different results when the two groups of examinees differ substantially on the anchor test. However, given that both methods rely on untestable assumptions, it is difficult to conclude which of the two equating approaches is more appropriate in a given situation. CE and PSE methods were compared from several perspectives in von Davier, Holland, and Thayer (2003, 2004a, 2004b) and von Davier (2003). These studies show that both PSE and CE appear to be similar in their standard errors of equating and in their degrees of population invariance. Thus, such *theoretical* considerations do not lead to a clear choice between the methods. von Davier et al. (2004a) gave an example where the two methods produce results that are sufficiently and reliably different enough to have practical consequences.

However, there is a series of empirical and simulation studies that have repeatedly found that, when there are large differences between the two groups in the NEAT design, the CE methods tend to show less bias and about the same variability as the corresponding PSE methods. Livingston, Dorans, and Wright (1990) observed this effect and they offered an explanation as to why the PSE methods are increasingly biased as the differences between the two groups on the anchor test increases. Livingston (2004) also discussed this phenomenon as did Dorans, Liu, and Hammond (2005). Wang, Lee, Brennan, and Kolen (2006) focused directly on comparing CE and PSE and, through an IRT-based simulation study, again showed that frequency estimation (a PSE method) is more biased than the chained equipercentile method when the groups differ in ability. Thus, it is increasingly clear that CE methods are preferable to PSE methods when the groups differ widely on the anchor test. However, Wright and Dorans (1993) showed that this is not always the case using an approach similar to that of Livingston et al. (1990) but employing a method of selecting the groups of examinees that more closely approximated the assumptions of PSE than did Livingston et al. At this point it is fair to say that rather than merely having theoretical shortcomings (Kolen & Brennan, 2004, p. 146), CE methods are clear competitors with PSE methods. However, a full understanding of when each is appropriate continues to elude us.

To compare the results of different equating methods, the usual approach is to design a study where a *true* or *criterion* equating is available and then to investigate the closeness of the different methods to the criterion equating. Examples of such studies that use the NEAT design and compare both CE and PSE methods are von Davier et al. (2006) and Wang et al. (2006). This approach is direct and simple in conception, but it does not allow for any detail in the explanation of why one method is closer to the criterion than the others. For the NEAT design this is especially problematic because both PSE and CE make different untestable assumptions about data that are not available in practice. A natural question is how adequate are these different sets of assumptions?

The present study uses the data from von Davier et al. (2006), where there is a natural criterion equating, but in addition, data are available in this study that are not usually available in practice. These extra data allow us to evaluate the underlying assumptions of CE and PSE. Reports on the agreement of CE and PSE with the criterion equating are given in von Davier et al. In that study, both the traditional linear-interpolation-based and the Gaussian kernel-

smoothing-based equipercentile functions of both the PSE and CE approaches were close to the criterion equipercentile function, but the CE results were slightly closer. The present study investigates how the *assumptions* of CE and PSE reflect these earlier findings.

The special data set from von Davier et al. (2006) will be described in more detail in a later section. The study uses the item responses from actual examinees taking a real test that was given at two different test administrations. The actual item responses from the whole test are used to create scores for smaller, nonoperational *pseudo-tests*. Furthermore, by ignoring some of the scores on the pseudo-tests for examinees in the different test administrations this approach can mimic the data in a NEAT design. This *ignored data* can be used to evaluate the usually untestable assumptions of CE and PSE, as we do here.

This report is part of a series of examinations of the special pseudo-test data set. von Davier and Ricker (2006) used the same data to compare PSE and CE to the criterion equating while varying both the type of anchor test (internal or external) and its length. Here, we use the same data to investigate the influence of the length (and, therefore, the reliability) of the anchor tests on the accuracy of the PSE and CE predictions.

The analyses discussed in this paper used loglinear models for presmoothing and the kernel equating (KE) method for computing the equipercentile functions needed (von Davier et al., 2004a). The previous study of von Davier et al. (2006) showed (a) that the KE version of PSE gave results that were very similar to frequency-estimation method that used linear-interpolation to continuize and (b) that the KE version of CE gave results that were very similar to the method of chained equipercentile equating that used linear-interpolation to continuize. For this reason, we did not think it necessary to use methods based on linear-interpolation continuization in our comparison of CE and PSE.

Basic Notation

In the NEAT design, the two operational tests to be equated, $\boldsymbol{X}$ and $\boldsymbol{Y}$, are given to two samples of examines from different test populations or administrations (denoted here by $\boldsymbol{P}$ and $\boldsymbol{Q}$). In addition, an anchor test, $\boldsymbol{A}$, is given to both samples from $\boldsymbol{P}$ and $\boldsymbol{Q}$. The data for the NEAT design is described in the *design table* (von Davier et al., 2004a) illustrated in Table 1, where ✓ denotes the presence of data and a blank indicates the absence of data.

Table 1

The Design Table for the NEAT Design

	X	***A***	***Y***
P	✓	✓	
Q		✓	✓

The anchor test score, ***A***, can be either a part of both ***X*** and ***Y*** (an *internal* anchor test) or a separate score that is not used for scoring the test (an *external* anchor test). Both types are considered in this study.

The *target population*, ***T***, for the NEAT design is the synthetic population based on ***P*** and ***Q,*** Braun and Holland (1982). ***P*** and ***Q*** are given weights that sum to 1, which denote their degree of influence on ***T***. Following Braun and Holland (1982), this is denoted by

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

If $w = 1$, then $\boldsymbol{T} = \boldsymbol{P}$ and if $w = 0$ then $\boldsymbol{T} = \boldsymbol{Q}$. If $w = ½$, then ***P*** and ***Q*** are represented equally in ***T***. Any choice of w between 0 and 1 is possible, and this reflects the amount of weight that is given to ***P*** and ***Q***. Other authors have indicated the *total populatio*n by summing the samples from ***P*** and ***Q***, (***P*** + ***Q***). This corresponds to taking w in (1) to be proportional to the sample size from ***P*** relative to the total for ***P*** and ***Q***, and this is the choice of w used here.

The pseudo-test data can be related to the design table in Table 1. ***P*** and ***Q*** correspond to the two test administrations that took the original basic test. The target population is the combined two test administrations and so corresponds to choosing w proportional to the sample size in administration ***P***. The pseudo-tests, ***X*** and ***Y*** and ***A***, are all formed from the real test items, with ***X*** and ***Y*** designed to be substantially different in difficulty and yet parallel in test content. Various types of pseudo-anchor tests were made to play the role of ***A***. Table 2 shows the design table for the pseudo-test data.

Table 2

The Design Table for the Pseudo-Test Data

	X	***A***	***Y***
P	✓	✓	✓
Q	✓	✓	✓

The pseudo-test data are then used to simulate a NEAT design by pretending that ***X*** was not given to ***Q*** and that ***Y*** was not given to ***P*** even though that was the case. The criterion equating used in von Davier et al. (2006) is found by using all the data from ***T*** for both ***X*** and ***Y*** and equating them through a single group design on ***T***. This criterion equating is the equating function that both CE and PSE attempt to estimate by making different types of assumptions about the missing data in Table 3. Thus, it is the natural criterion equating for the pseudo-test data.

In our discussion we will let *F*, *G,* and *H* denote the cumulative distribution functions (cdfs) of ***X***, ***Y,*** and ***A***, respectively, and will further specify the populations on which these cdfs are determined by the subscripts ***P***, ***Q***, and ***T***. These cdfs arise throughout the following discussion.

All OSE methods may be viewed as based on the equipercentile equating function defined on the target population, T, as:

$$e_{XY;T}(x) = G_T^{-1}(F_T(x)) \qquad (2)$$

where $F_T(x)$ and $G_T(y)$ are the cdfs of ***X*** and ***Y***, respectively, on ***T***.

Linear equating may be derived from (2) by assuming that $F_T(x)$ and $G_T(y)$ are continuous and have the same shape with possibly differing means and variances. Under this assumption, the equipercentile equating function, $e_{XY;T}(x)$, reduces to the linear equating function, $Lin_{XY;T}(x)$, defined by

$$Lin_{XY;T}(x) = \mu_{YT} + \sigma_{YT}((x - \mu_{XT})/\sigma_{XT}). \qquad (3)$$

Equating Methods for the NEAT Design

In the NEAT design, the two operational tests, ***X*** and ***Y***, are each observed *either on* ***P*** *or on* ***Q***, *but not both*. Thus, ***X*** and ***Y*** are not *both* observed on ***T***, regardless of the choice of *w*. For this reason, assumptions must be made to overcome the missing data that arise in the NEAT design, which are evident in Table 1. A basic task for developing OSE methods for the NEAT design is to make acceptable and sufficiently strong assumptions that allow values for $F_T(x)$ and $G_T(y)$ to be found. In other equating and test linking designs, such as the equivalent-groups or the single-group designs, the target population is simply the group from which the examinees were sampled. In those cases, $F_T(x)$ and $G_T(y)$ may be directly estimated from the observed data. In the

NEAT design, however, assumptions that are not directly testable using the available data must be added to the mix. CE and PSE represent two different sets of such assumptions.

In the NEAT design, OSEs have been proposed that use the anchor test information in three fundamentally different ways. First, the anchor score can be used as a stratifying or conditioning variable for estimating the score distributions or the sample statistics of the tests to be equated. This approach is similar to poststratification in survey research, and, following von Davier et al. (2004a), we refer to equating methods based on this approach as *poststratification equating* (PSE) methods. Second, the anchor score can be used as the middle link in a chain of linking relationships; $\boldsymbol{X}$ is first linked to $\boldsymbol{A}$ and then $\boldsymbol{A}$ is linked to $\boldsymbol{Y}$. Following standard usage, we will refer to equating methods based on this approach as *chain equating* (CE) methods. The equating functions for either PSE or CE can be linear or nonlinear in shape.

A third way to use the anchor information uses classical test theory to produce estimates of the mean and variance of $\boldsymbol{X}$ and $\boldsymbol{Y}$ over $\boldsymbol{T}$. This results in the Levine OSE linear method. Currently, there is no equipercentile version of the Levine OSE linear method and therefore we do not consider this third approach in this paper.

The two OSE methods, CE and PSE, make different assumptions about the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the populations where they are *not observed*. These assumptions were identified in von Davier et al. (2004a) and are briefly given here.

> CE assumptions. The equipercentile function computed on $\boldsymbol{P}$ for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ is the same as that for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ on $\boldsymbol{T}$ for any choice of $\boldsymbol{T} = w\boldsymbol{P} + (1 - w)\boldsymbol{Q}$. An analogous assumption holds for the links from $\boldsymbol{A}$ to $\boldsymbol{Y}$ in $\boldsymbol{Q}$ and in $\boldsymbol{T}$.
>
> PSE assumptions. The conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{P}$ is the same as the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{T}$, for any choice of $\boldsymbol{T} = w\boldsymbol{P} + (1 - w)\boldsymbol{Q}$. An analogous assumption holds for $\boldsymbol{Y}$ given $\boldsymbol{A}$ in $\boldsymbol{Q}$ and in $\boldsymbol{T}$.

In the pseudo-test data we can check the CE and PSE assumptions. However, as given above, these assumptions require checking different things for CE and for PSE. For example, checking the CE assumptions requires comparing the link from $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{P}$ with the link from $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{T}$, and similarly, for the links between $\boldsymbol{A}$ and $\boldsymbol{Y}$ in $\boldsymbol{Q}$ and $\boldsymbol{T}$. However, checking the PSE assumptions requires comparing the conditional distributions of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{P}$ with the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{T}$, and similarly for the conditional distribution of $\boldsymbol{Y}$

given $\boldsymbol{A}$ in $\boldsymbol{Q}$ and $\boldsymbol{T}$. Because the pseudo-test data includes data for both $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$, these checks of the (usually) untestable assumptions are possible, at least in principle.

One problem that immediately arises is that what is compared in the evaluation of the CE assumptions is very different from what is compared in the evaluation of the PSE assumptions. It is not clear how to compare the magnitude of a failure of the CE assumptions with the magnitude of a failure of the PSE assumptions. Our solution to this problem is to identify necessary consequences of the two sets of assumptions for the distribution of $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$ and then to compare these consequences or *predictions* with the actual data. We discuss these predictions next.

The Predictions of PSE and CE

The PSE predictions. The PSE assumptions are supposed to hold for any choice of $\boldsymbol{T} = w\boldsymbol{P} + (1 - w)\boldsymbol{Q}$, so that in particular they hold for $\boldsymbol{T} = \boldsymbol{Q}$. Thus, the PSE assumptions imply that the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{A}$ in $\boldsymbol{T}$ may be expressed as

$$\mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{A}, \boldsymbol{Q}\} = \mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{A}, \boldsymbol{P}\}. \tag{4}$$

Hence, the marginal distribution of $\boldsymbol{X}$ in $\boldsymbol{Q}$, $f_{x\boldsymbol{Q}} = \mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{Q}\}$, is given by

$$f_{x\boldsymbol{Q}} = \mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{Q}\} = \sum_{a} \mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{A} = a, \boldsymbol{P}\} h_{a\boldsymbol{Q}}, \tag{5}$$

where

$$h_{a\boldsymbol{Q}} = \mathrm{P}\{\boldsymbol{A} = a \mid \boldsymbol{Q}\}, \tag{6}$$

is the marginal distribution of $\boldsymbol{A}$ in $\boldsymbol{Q}$. Thus, (5) is the PSE prediction of $f_{x\boldsymbol{Q}}$ that is a necessary consequence of the PSE assumptions. Similar predictions for the marginal distribution of $\boldsymbol{Y}$ in $\boldsymbol{P}$ follow from the PSE assumption for the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{A}$ in $\boldsymbol{T}$ and $\boldsymbol{P}$.

Because the PSE predictions are necessary consequences of the PSE assumptions, any evidence that the PSE predictions are wrong implies that the PSE assumptions are also wrong. To implement these PSE predictions we used the following approach.

First, we used loglinear models to presmooth the bivariate distribution of ($\boldsymbol{X}$, $\boldsymbol{A}$) obtained from $\boldsymbol{P}$ and the bivariate distribution of ($\boldsymbol{Y}$, $\boldsymbol{A}$) from $\boldsymbol{Q}$. We denoted these presmoothed bivariate probabilities, respectively, as

$$p_{xa} = \mathrm{P}\{\boldsymbol{X} = x, \boldsymbol{A} = a \mid \boldsymbol{P}\} \text{ and } q_{ya} = \mathrm{P}\{\boldsymbol{Y} = y, \boldsymbol{A} = a \mid \boldsymbol{Q}\}. \quad (7)$$

Second, using these bivariate probabilities, form the marginal distributions of $\boldsymbol{A}$ in $\boldsymbol{P}$ and $\boldsymbol{Q}$, that is,

$$h_{a\boldsymbol{P}} = \sum_x p_{xa} \text{ and } h_{a\boldsymbol{Q}} = \sum_y q_{ya}. \quad (8)$$

Third, we computed the conditional probability, $\mathrm{P}\{\boldsymbol{X} = x \mid \boldsymbol{A} = a, \boldsymbol{P}\}$, as the ratio $p_{xa}/h_{a\boldsymbol{P}}$.

Fourth, we used these estimated conditional probabilities to obtain the predicted score probabilities for $\boldsymbol{X}$ in $\boldsymbol{Q}$ via equation (5), that is,

$$f_{x\boldsymbol{Q}} = \sum_a p_{xa}(h_{a\boldsymbol{Q}}/h_{a\boldsymbol{P}}). \quad (9)$$

By similar reasoning, the predicted score probabilities for $\boldsymbol{Y}$ in $\boldsymbol{Q}$ are

$$g_{y\boldsymbol{P}} = \sum_a q_{ya}(h_{a\boldsymbol{P}}/h_{a\boldsymbol{Q}}). \quad (10)$$

We denoted the observed frequencies of $\boldsymbol{X}$ in $\boldsymbol{Q}$ by $n_{x\boldsymbol{Q}}$ and the frequencies for $\boldsymbol{Y}$ in $\boldsymbol{P}$ by $m_{y\boldsymbol{P}}$. In a real NEAT design, neither of these two sets of frequencies is available, but in this special data set they are. The $\boldsymbol{X}$ in $\boldsymbol{Q}$ frequencies, $\{n_{x\boldsymbol{Q}}\}$, sum to $N_{\boldsymbol{Q}}$, while the $\boldsymbol{Y}$ in $\boldsymbol{P}$ frequencies sum to $N_{\boldsymbol{P}}$. The check on the assumptions of PSE that we propose is to compare the *predicted frequencies*, $N_{\boldsymbol{Q}} f_{x\boldsymbol{Q}}$ and $N_{\boldsymbol{P}} g_{y\boldsymbol{P}}$, to the *observed frequencies*, $n_{x\boldsymbol{Q}}$ and $m_{y\boldsymbol{P}}$, respectively. We compare the observed with the predicted frequencies in several ways described in more detail, later. These include direct comparisons of the observed and predicted frequencies as well as the *smoother* comparisons of observed and predicted moments.

The CE predictions. The CE assumptions do not directly concern discrete score distributions as the PSE assumptions do. Instead, they assert that the equipercentile function for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{T}$ is the same for any choice of $\boldsymbol{T}$, including both $\boldsymbol{P}$ and $\boldsymbol{Q}$. The CE predictions for the score distributions of $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$ are not as direct as they are for PSE. We used the following approach.

First, we took the presmoothed bivariate distributions, $\{p_{xa}\}$ and $\{q_{ya}\}$, from (7) and used them to get the marginal score probabilities of $\boldsymbol{X}$ in $\boldsymbol{P}$, $\boldsymbol{A}$ in $\boldsymbol{P}$, $\boldsymbol{A}$ in $\boldsymbol{Q}$, and $\boldsymbol{Y}$ in $\boldsymbol{Q}$, denoted, respectively, by $f_{x\boldsymbol{P}}$, $h_{a\boldsymbol{P}}$, $h_{a\boldsymbol{Q}}$, and $g_{y\boldsymbol{Q}}$, in parallel with the notation in (8) – (10).

Second, using the methods of kernel equating, we continuized these score probabilities to get the continuous cdfs $F_{\boldsymbol{P}}(x)$, $H_{\boldsymbol{P}}(a)$, $H_{\boldsymbol{Q}}(a)$, and $G_{\boldsymbol{Q}}(y)$.

Third, we observed that the assumption that the equipercentile function for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{P}$ is the same as it is for linking $\boldsymbol{X}$ to $\boldsymbol{A}$ in $\boldsymbol{Q}$ means that $H^{-1}_{\boldsymbol{Q}}(F_{\boldsymbol{Q}}(x)) = H^{-1}_{\boldsymbol{P}}(F_{\boldsymbol{P}}(x))$, so that the CE predicted *continuized* cdf of $\boldsymbol{X}$ in $\boldsymbol{Q}$ is given by

$$F_{\boldsymbol{Q}}(x) = H_{\boldsymbol{Q}}(H^{-1}_{\boldsymbol{P}}(F_{\boldsymbol{P}}(x))). \tag{11}$$

Similarly, the CE predicted continuized cdf of $\boldsymbol{Y}$ in $\boldsymbol{P}$ is

$$G_{\boldsymbol{P}}(y) = H_{\boldsymbol{P}}(H^{-1}_{\boldsymbol{Q}}(G_{\boldsymbol{Q}}(y))). \tag{12}$$

To compute $F_{\boldsymbol{Q}}(x)$ in (11) for any x, first compute $a = e_{\boldsymbol{A}}(x) = H^{-1}_{\boldsymbol{P}}(F_{\boldsymbol{P}}(x))$, the KE equipercentile function linking $\boldsymbol{X}$ to $\boldsymbol{A}$ on $\boldsymbol{P}$. Then, using the computed value of a, compute $H_{\boldsymbol{Q}}(a)$ from the KE continuized cdf. The new KE software (ETS, 2005) has subroutines for both of these calculations. Similar calculations are made for $G_{\boldsymbol{P}}(y)$ in (12).

The two predicted cdfs are continuous, but the score data are *discrete*. In order make comparisons with the PSE predictions, we suggest *discretizing* the two predicted continuous cdfs in the following way. Denote the $\boldsymbol{X}$-scores by x_j, for $j = 1$ to J and evaluate $F_{\boldsymbol{Q}}(x)$ at the value, $x = (x_j + x_{j+1})/2$, for $j = 1$ to $J - 1$. Then, define a discrete probability distribution, $\{r_{j\boldsymbol{Q}}\}$, for $\boldsymbol{X}$ by

$$r_{j\boldsymbol{Q}} = F_{\boldsymbol{Q}}((x_j + x_{j+1})/2) - F_{\boldsymbol{Q}}((x_{j-1} + x_j)/2), \text{ for } j = 2, 3, \ldots, J - 1,$$

$$r_{J\boldsymbol{Q}} = 1 - F_{\boldsymbol{Q}}((x_{J-1} + x_J)/2), \text{ and } r_{1\boldsymbol{Q}} = F_{\boldsymbol{Q}}((x_1 + x_2)/2). \tag{13}$$

The $\{r_{j\boldsymbol{Q}}\}$, given in (13), are discrete probabilities that sum to 1.0 and that, if continuized using the scores, $\{x_j\}$, will closely reproduce the cdf, $F_{\boldsymbol{Q}}(x)$. This discretization of the continuous cdf, $F_{\boldsymbol{Q}}(x)$, is of possible independent interest and allows the score points to be unequally spaced,

when this occurs. However, in the present study, number-right scores arise and these are equally spaced. We propose another use for this method of discretizing cdfs in the discussion section.

In the same way, the predicted score probabilities for $\boldsymbol{Y} = y_k$ (k = 1 to K) in $\boldsymbol{P}$ are given by

$$s_{k\boldsymbol{P}} = G_{\boldsymbol{P}}((y_k + y_{k+1})/2) - G_{\boldsymbol{P}}((y_{k-1} + y_k)/2), \text{ for } k = 2, 3, \ldots, K-1, \text{ and} \tag{14}$$

$$s_{K\boldsymbol{P}} = 1 - G_{\boldsymbol{P}}((y_{K-1} + y_K)/2), \text{ and } s_{1\boldsymbol{P}} = G_{\boldsymbol{P}}((y_1 + y_2)/2).$$

We used the values of $N_{\boldsymbol{Q}}\, r_{j\boldsymbol{Q}}$ as the CE prediction of the $\boldsymbol{X}$ in $\boldsymbol{Q}$ frequencies, $n_{x\boldsymbol{Q}}$, just as $N_{\boldsymbol{Q}}\, f_{x\boldsymbol{Q}}$ is the PSE prediction of these frequencies. In a similar way, we used $N_{\boldsymbol{P}}\, s_{k\boldsymbol{P}}$ as the CE prediction of the $\boldsymbol{Y}$ in $\boldsymbol{P}$ frequencies, $m_{y\boldsymbol{P}}$.

Except for the discretizing steps, (13) and (14), that are needed to make the CE predictions comparable to those of PSE, the CE predictions are necessary consequences of the CE assumptions. Hence, evidence that they are wrong is evidence that the CE assumptions are wrong.

*Comparing the predicted frequencies with the data***.** There are several different ways to investigate the difference between any set of *observed* and *predicted* score frequencies. We use three different approaches in our analyses.

First, to get an overall view of how well the predictions tracked the observed frequencies we *graphed* the observed and predicted frequencies together as well as their Freeman-Tukey (FT) residuals (Holland & Thayer, 2000) to display the full set of predicted and observed frequencies. The FT residuals have the form,

$$\sqrt{n_i} + \sqrt{n_i + 1} - \sqrt{4m_i + 1}, \tag{15}$$

where n_i denotes the observed frequencies and m_i the predicted frequencies for either CE or PSE. If the observed frequencies are well approximated by the predictions, then these residuals will tend to show no pattern and to lie in the range expected for approximate normal deviates, that is, plus or minus 2 or 3.

Second, to get a more quantitative and summary assessment of the agreement between the observed and predicted frequencies we used three standard goodness-of-fit measures—

likelihood ratio chi-square, Pearson chi-square, and sum of squared FT residuals (Holland & Thayer, 2000). The following formulas define these measures. In each case, n_i denotes the observed frequencies and m_i the corresponding predicted frequencies from either CE or PSE.

$$\text{Pearson } \chi^2 \text{ statistic, } \chi^2 = \sum_i \frac{(n_i - m_i)^2}{m_i}, \tag{16}$$

$$\text{Likelihood ratio } \chi^2 \text{ statistic, } G^2 = 2\sum_i n_i \log(n_i / m_i), \tag{17}$$

$$\text{The FT } \chi^2 \text{ statistic, } \chi^2_{FT} = \sum_i (\sqrt{n_i} + \sqrt{n_i + 1} - \sqrt{4m_i + 1})^2 . \tag{18}$$

These three measures are often used to assess the closeness of fitted frequencies to observed frequencies in discrete distributions of scores (Holland & Thayer, 2000) and have nominal chi-square reference distributions when the disagreement between the observed and predicted frequencies is due to random variation. However, in this application these reference distributions are not likely to be accurate because none of these predicted frequencies were created under the assumptions that would lead to using these references distributions. Nonetheless, the measures are useful quantitative and summary indices of the overall agreement between the observed and predicted frequencies. Two reviewers suggested comparing the predicted frequencies to a set of *smoothed* versions of the observed frequencies (rather than the raw unsmoothed observed frequencies) to dampen some of the noise in the observed frequencies. However, we resisted this suggestion since it would introduce the *method of smoothing the raw frequencies* into the comparison and, in our opinion, would cloud rather than simplify it. Moreover, our third type of comparison does provide a natural type of smoothing of the observed frequencies to compare with the predicted frequencies, to which we now turn.

For a smoother and more detailed look at the predictions, we compared the first four moments—mean, standard deviation, skewness, and kurtosis—of the predicted and observed frequencies and used the percent relative difference between the observed and predicted moments as a way to quantify the relative accuracy of the predictions.

Study Details

The original data set. The data come from one test form of a licensing test program for prospective teachers of children in primary through upper elementary school grades. The form included 119 multiple-choice items, about equally divided among four content areas—language arts, mathematics, social studies, and science. This form of the test was administered twice, and the two test administrations play the role of populations ***P*** and ***Q*** in our analysis. The mean total scores (number right) of the examinees taking the test at these two administrations differed by approximately one-fourth of a standard deviation, as can be seen in Table 3. This data set was selected because of the large number of test items from which to construct pseudo-tests and because the score distributions at the two test administrations were substantially different.

Table 3

Ns, Means, and Standard Deviations of the Original Test Score for the Two Test Administrations, P and Q

Administration	***P***	***Q***
Number of examinees	6,168	4,237
Mean score	82.3	86.2
SD of scores	16.0	14.2

Test construction. We used these data to construct two *pseudo-tests*, ***X*** and ***Y***, as well as three different *pseudo-anchor tests*, ***A*1**, ***A*2,** and ***A*3**, of different lengths. A pseudo-test consists of a subset of the test items from the original 119-item test, and the score on the pseudo-test for an examinee in the sample is found from the item responses of that examinee to the items in the pseudo-test. This approach is an alternative to simulating test data from an item response model and has the benefit of being based on real test data from real examinees rather than being completely based on a statistical model.

The external anchor test cases. To create data sets with *external anchor tests*, we used the 119 items from the original test to create two smaller pseudo-tests, ***X*** and ***Y***. Each of these contained 44 items, 11 items from each of the four content areas. Care was given to make ***X*** and ***Y*** parallel in content but different in difficulty. Test ***X*** was constructed to be easier than ***Y***, based on the item statistics for the items. Tests ***X*** and ***Y*** had no items in common. In addition, a basic

set of 24 items (6 from each content area) was selected to be representative of the original test and to serve as the largest external anchor, ***A*1**. The two other anchor tests, ***A*2** and ***A*3**, were formed by deleting 4 and 8 items, respectively, from ***A*1** in such a way that ***A*2** is a 20-item subset of ***A*1**, and ***A*3** is a 16-item subset of ***A*1** and ***A*2**. Furthermore, to maintain parallelism in content, test ***A*2** had five items from each content area, while ***A*3** had four.

The pseudo-anchor tests were constructed (to the extent possible) to cover the content tested by the 119-item test and the two 44-items tests as well as to represent the content categories in the same proportions as the original test. The mean difficulty of the anchor tests approximately equaled the mean for the original test. The structure of the various pseudo-tests is outlined in Table 4.

Table 4

The Structure of the Two Basic Pseudo-Tests and the Three External Anchor Tests

X The easier test	Anchor items	***Y*** The more difficult test
	Language arts	
1, 5, 6, 7, 8, 9, 11, 23, 24, 25, 30	***A*1**: 3, 10, 14, 15, 17, 18 ***A*2**: 3, 10, 14, 17, 18 ***A*3**: 3, 10, 14, 18	2, 4, 12, 13, 19, 20, 21, 26, 27, 28, 29
	Mathematics	
31, 33, 34, 40, 44, 46, 47, 49, 51, 54, 60	***A*1**: 32, 42, 43, 52, 55, 58 ***A*2**: 42, 43, 52, 55, 58 ***A*3**: 42, 43, 52, 58	35, 37, 38, 41, 45, 48, 50, 53, 56, 57, 59
	Social studies	
61, 63, 66, 67, 69, 77, 78, 83, 86, 87, 90	***A*1**: 64, 71, 73, 74, 76, 79 ***A*2**: 64, 71, 74, 76, 79 ***A*3**: 64, 71, 74, 79	62, 65, 68, 70, 72, 75, 80, 81, 82, 85, 88
	Science	
92, 93, 95, 99, 103, 105, 106, 108, 113, 114, 118	***A*1**: 91, 98, 101, 107, 110, 120 ***A*2**: 91, 98, 101, 110, 120 ***A*3**: 91, 98, 101, 110	94, 96, 97, 100, 102, 104, 109, 112, 115, 116, 117

Note. Item numbers are from the original test.

From Table 4 it is clear that the original test and the pseudo-tests, ***X***, ***Y***, ***A1***, ***A2,*** and ***A3,*** are *multidimensional* in the sense that the content covered involved four topic areas. However, great care was taken to make all of the pseudo-tests as parallel in this content coverage as possible, and each test has proportionally the same content coverage across the four dimensions. Hence, to the extent possible, all of the pseudo-tests are multidimensional in the *same* way.

Table 5 gives the *N*s, means, standard deviations, and alpha reliabilities of the scores on ***X***, ***Y***, ***A1***, ***A2,*** and ***A3*** and for the two sums ***X1*** = ***X*** + ***A1*** and ***Y1*** = ***Y*** + ***A1*** (they play a role for the internal anchor cases, see below) for the examinees in ***P***, ***Q,*** and the combined group. ***X*** is easier than ***Y*** because the mean score on ***X*** is higher than the mean score on ***Y*** in all three groups. For example, the mean score on ***X*** on the combined group is larger than the mean score on ***Y*** by approximately 127% of the ***Y***-standard deviation. This difference in difficulty is substantial and is probably an extreme that would only be observed in practice if pretesting is not feasible. In addition, all three anchor tests show approximately 23 to 24% difference between ***P*** and ***Q***, in terms of the standard deviation of the combined group. The reliabilities of the three anchor tests behave as expected, with ***A1*** the most reliable and ***A3*** the least reliable. However, the range of these reliabilities is not large—from .68 to .75 on the combined group.

Table 5

Ns, Means, (Standard Deviations), and [Alpha Reliabilities] of the Scores on X, Y A1, A2, A3, X1, and Y1 in P, Q, and the Combined Group, P + Q

Test	***X***	***Y***	***A1***	***A2***	***A3***	***X1*** = ***X*** + ***A1***	***Y1*** = ***Y*** + ***A1***
P	35.1	26.6	16.0	13.7	10.8	51.2	42.6
	(5.7)	(6.7)	(4.2)	(3.6)	(3.0)	(9.3)	(10.3)
N = 6,168	[.81]	[.81]	[.75]	[.71]	[.68]	[.88]	[.88]
Q	36.4	28.0	17.0	14.5	11.5	53.4	45.0
	(4.8)	(6.3)	(3.9)	(3.3)	(2.8)	(8.0)	(9.6)
N = 4,237	[.77]	[.79]	[.73]	[.69]	[.66]	[.85]	[.87]
P + ***Q***	35.6	27.2	16.4	14.0	11.1	52.1	43.6
	(5.4)	(6.6)	(4.1)	(3.5)	(3.0)	(8.9)	(10.1)
N = 10,405	[.80]	[.80]	[.75]	[.71]	[.68]	[.87]	[.87]

The internal anchor test cases. To create data sets that had *internal anchor tests*, we formed ***X*1** = ***X*** + ***A*1** and ***Y*1** = ***Y*** + ***A*1**. Then we paired ***X*1** and ***Y*1** with ***A*1**, ***A*2,** or ***A*3** as the three internal anchor test scores. Because ***A*2** was a subset of ***A*1** and ***A*3** was a subset of ***A*1** and ***A*2**, each of the three anchor tests is internal to the total scores, ***X*1** and ***Y*1**. This approach allowed us to keep the total test the same size (44 + 24 = 68 items) as we varied the length (and therefore the reliabilities) of the anchor tests.

Mimicking the NEAT design. Because all the examinees in ***P*** and ***Q*** took all the 119 items on the original test, it follows that all of the examinees in ***P*** and ***Q*** also have scores for the two 44-item tests, ***X*** and ***Y***, as well as for each of the three anchor tests, ***A*1**, ***A*2,** and ***A*3**. In order to mimic the structure of the NEAT design indicated in Table 1, we pretended scores for ***X*** or ***X*1** were not available for the examinees in the test administration designated as ***Q*** and that scores for ***Y*** or ***Y*1** were not available for the examinees in ***P***. Thus, the data indicated in Table 2 can be viewed as the NEAT design in Table 1. However, because all scores were, in fact, available, they allow us to test the different assumptions made by CE and PSE in the NEAT design using the predictions discussed in the earlier sections.

Presmoothing the bivariate score distributions. The same polynomial loglinear model was used for presmoothing all of the bivariate score distributions that arose from the joint frequency distributions of a pseudo-test and an anchor test. Appropriate adjustments were made for the structural zeros in the case of the internal anchor tests. The model was selected after considerable analysis of the various bivariate distributions using a variety of possible loglinear models. The chosen bivariate model fit five marginal moments for each score variable plus four cross-product moments of the form xa, xa^2, x^2a, and x^2a^2. Examination of the marginal and conditional distributions of the bivariate frequencies indicated that a loglinear model of this form fit all the sample bivariate distributions well.

Continuizing the cdfs. All of the cdfs were continuized using Gaussian kernel smoothing with a penalty function that minimized the sum of squared discrepancies between the presmoothed probabilities and density function of the final cdfs (von Davier et al., 2004a). The resulting data-dependent bandwidths ranged from 0.529 to 0.640. These values are typical of those obtained for presmoothed data where the loglinear models are of the polynomial forms described earlier.

Results

This section focuses on comparisons of the predictions made by CE and PSE with the observed data for ***X*** or ***X*1** in ***Q*** and for ***Y*** or ***Y*1** in ***P***. The results are divided into three parts. First, graphs of the observed and predicted frequencies are examined to assess the overall agreement or disagreement between the predicted and observed data. Second, more quantitative assessments of the agreement between the observed and predicted frequencies are made using three different measures of goodness-of-fit between the observed and predicted distributions. Third, we give detailed comparisons of the first four moments of the observed and predicted distributions.

Comparisons of the observed and predicted frequencies. Figures 1 and 2 graph the observed and predicted frequencies for CE and PSE for ***X*** and ***X*1** in ***Q*** and for ***Y*** and ***Y*1** in ***P***, for the case of the longest anchor test, ***A*1**. (All of the graphs for the shorter anchor tests look very similar and are given in the appendix.)

It is evident that the predictions of CE and PSE are very similar and that notable departures of the observed frequencies from CE are associated with notable departures from PSE as well. To look at the differences in more detail, we use the Freeman-Tukey residuals that are graphed in Figures 3 and 4.

Examination of Figures 3 and 4 indicates that the pattern of the residuals for CE and PSE are very similar and that it appears fairly random, well within the expected range for well-fitting predictions. However, those for CE often are smaller than those for PSE. This is clearest in the middle range of scores in Figure 3. In summary, the graphical plot of the predictions of CE and PSE show that they both track the data fairly well and both sets of predictions appear to be somewhat more similar to each other than they are to the observed data. The next comparisons looks at the overall agreement of the predictions in a more summary and quantitative way.

Comparisons of the goodness-of-fit measures. Table 6 gives the values for χ^2, G^2, and χ^2_{FT}, defined earlier, for all the cases in the study.

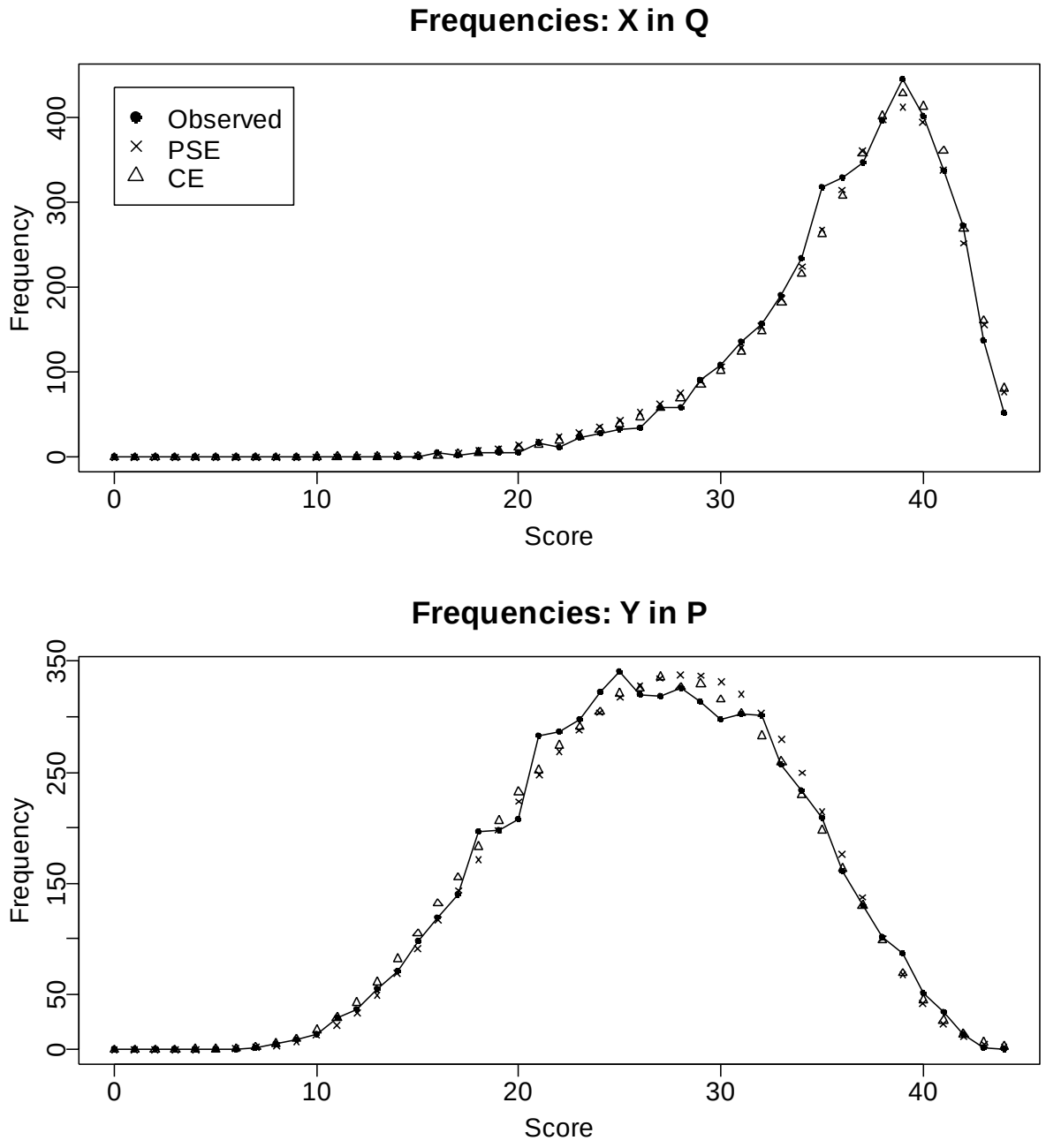

Figure 1. **Frequencies for *X* in *Q* and *Y* in *P* for external anchor test *A*1.**

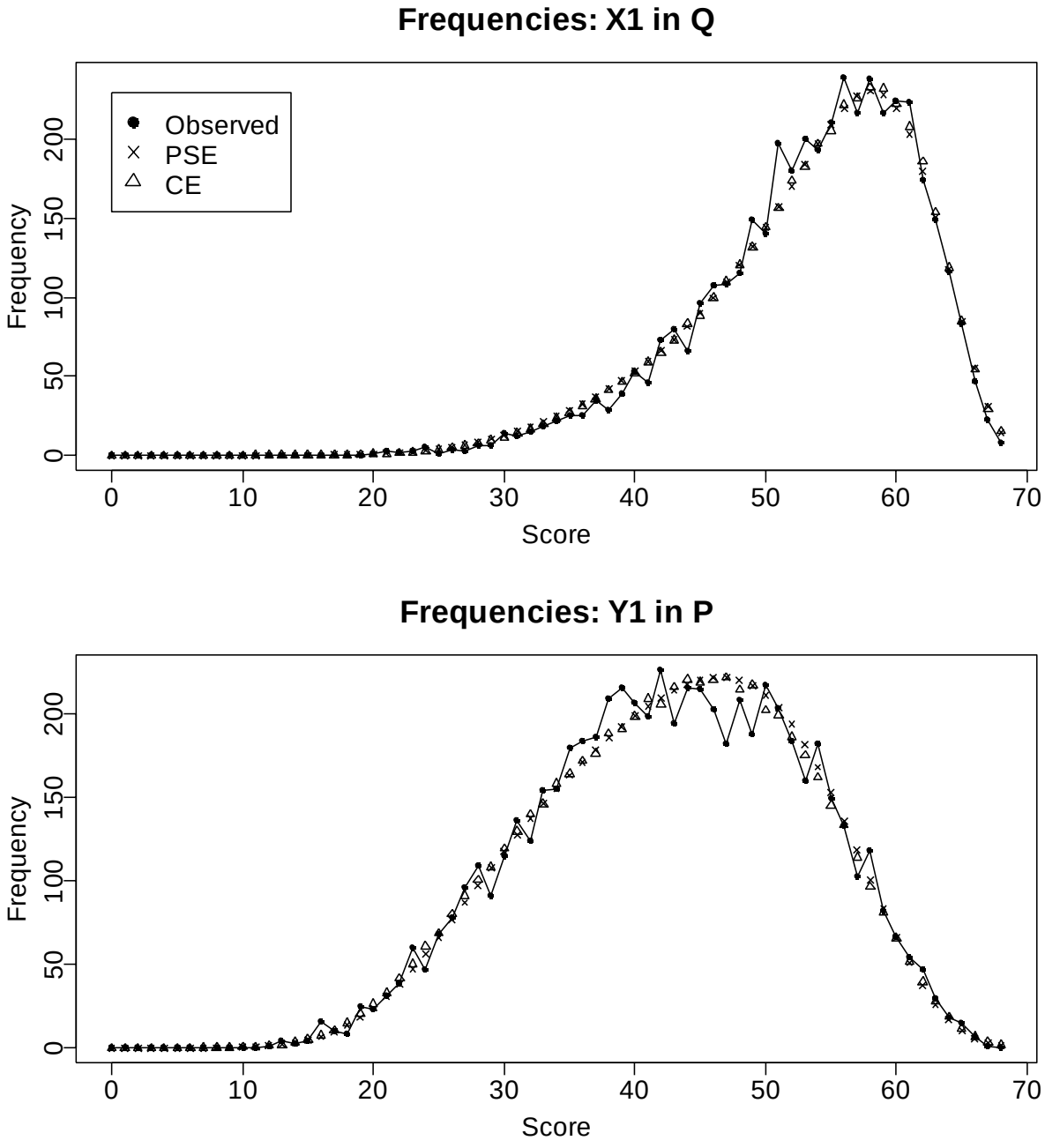

Figure 2. **Frequencies for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*1.**

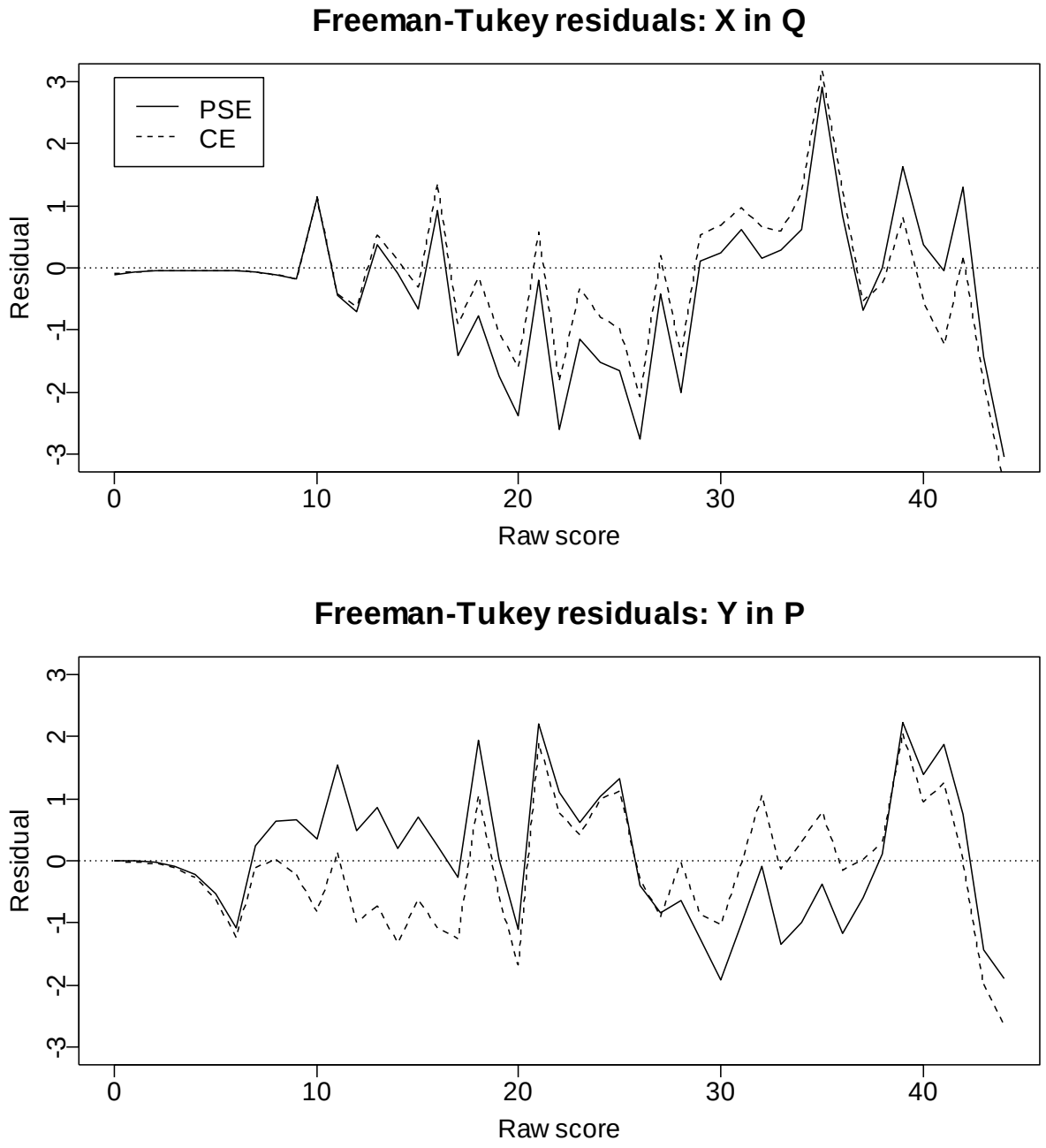

Figure 3. **Freeman-Tukey residuals for *X* in *Q* and *Y* in *P* for external anchor test *A*1.**

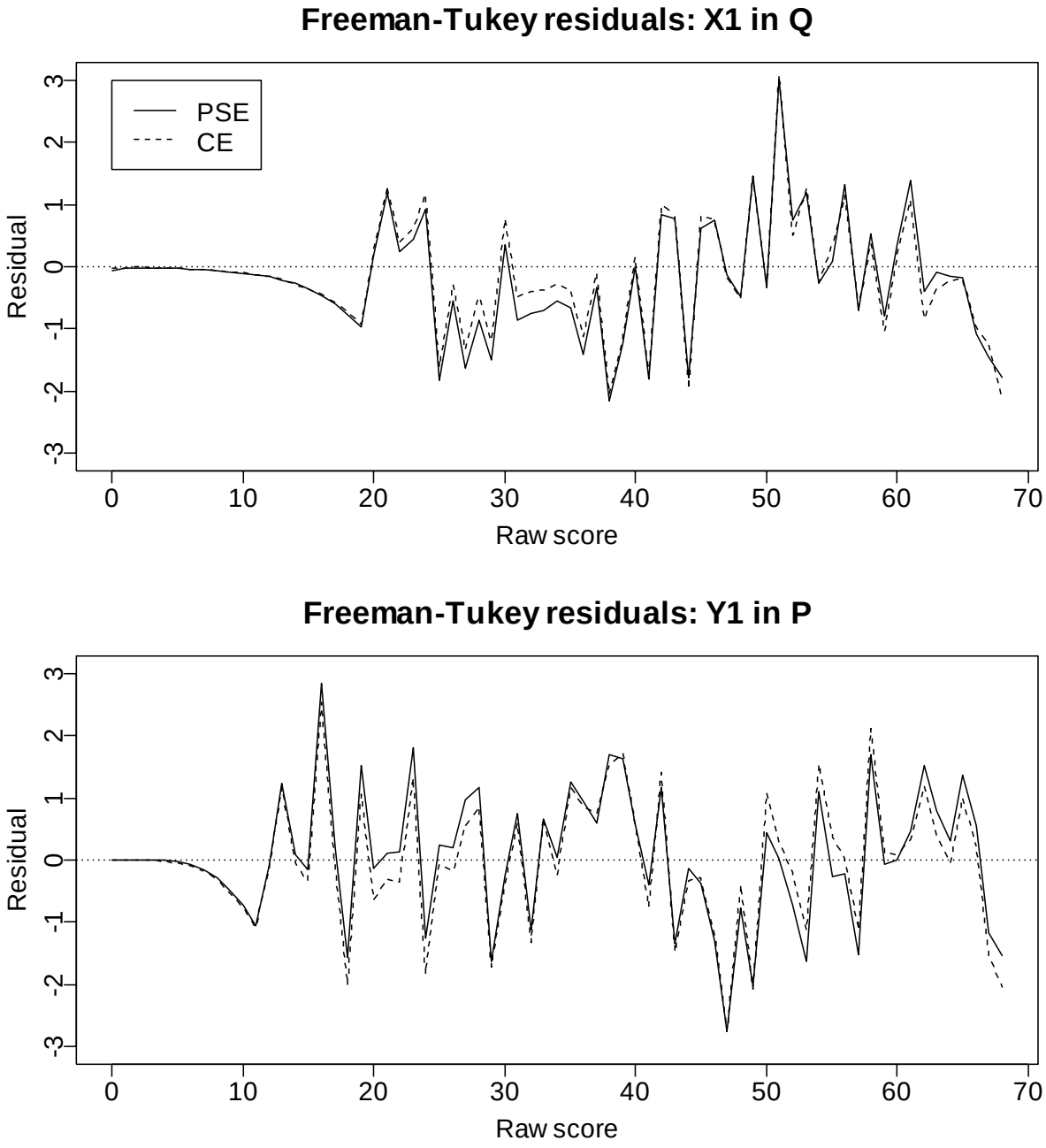

Figure 4. **Freeman-Tukey residuals for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*1.**

Table 6

The Three Goodness-of-Fit Measures

External anchor cases				Internal anchor cases			
Tests and anchors	χ^2	G^2	χ^2_{FT}	Tests and anchors	χ^2	G^2	χ^2_{FT}
X*, *A*1: *Q				***X*1, *A*1: *Q***			
PSE	63.9	68.5	66.3	PSE	58.6	64.2	59.5
CE	56.6	57.1	54.5	CE	56.4	60.5	56.0
***X*, *A*2**				***X*1, *A*2**			
PSE	72.0	77.6	76.0	PSE	69.1	76.1	71.8
CE	72.5	74.7	73.4	CE	67.0	72.4	68.4
***X*, *A*3**				***X*1, *A*3**			
PSE	78.0	86.7	85.1	PSE	75.5	85.0	79.8
CE	66.7	71.8	69.4	CE	71.8	78.9	73.6
Y*, *A*1: *P				***Y*1, *A*1: *P***			
PSE	49.3	51.4	49.9	PSE	78.1	77.7	74.1
CE	37.9	43.3	42.4	CE	74.8	77.9	74.9
***Y*, *A*2**				***Y*1, *A*2**			
PSE	58.7	59.8	58.0	PSE	90.2	88.1	84.5
CE	46.7	50.2	48.2	CE	82.4	84.1	81.1
***Y*, *A*3**				***Y*1, *A*3**			
PSE	68.3	68.9	67.4	PSE	103.7	97.9	94.2
CE	45.4	48.5	48.1	CE	91.6	89.6	87.2

The results in Table 6 quantify the observation made earlier from Figures 3 and 4 that the predictions of CE are somewhat closer to the observed frequencies than the PSE predictions. In all but three cases, all of the goodness-of-fit measures are smaller for CE than for PSE. In the three situations where this is not true (the shaded cells of Table 6) the difference between the goodness-of-fit measures for CE and PSE values is small. Thus, while the CE and PSE predictions are very similar, as seen in Figures 1 and 2, those of CE are usually slightly closer to the observed frequencies.

In addition, there is a consistent tendency for the goodness-of-fit measures for PSE to get smaller as the length of the anchor test increases. This trend is consistent in every case in Table 7.

Thus, it is evident that the length (and the reliability) of the anchor test has a distinct and measurable effect on improving the predictions of PSE. This effect is not easily seen in the graphs of the frequencies. The predictions of CE do not show this trend for the external anchor test cases but they do show it for the internal anchor test cases.

Table 7

The First Four Moments of the Observed and the Predicted Distributions and Their Relative Differences. External Anchor Test Cases

Tests, anchor; obs/pred.	Mean	Mean % rel. dif.	SD	SD % rel. dif.	Skewness	Skew % rel. dif.	Kurtosis	Kurt % rel. dif.
X, A*1: *Q								
Obs	36.38		4.77		-1.09		1.54	
PSE	36.16	-0.6	5.15	7.9	-1.09	-0.5	1.30	-15.6
CE	36.42	0.1	5.00	4.7	-1.11	-2.2	1.41	-8.4
***X, A*2**								
Obs	36.38		4.77		-1.09		1.54	
PSE	36.13	-0.7	5.19	8.8	-1.08	1.0	1.22	-20.9
CE	36.41	0.1	5.05	5.7	-1.08	0.8	1.28	-17.0
***X, A*3**								
Obs	36.38		4.77		-1.09		1.54	
PSE	36.04	-0.9	5.26	10.2	-1.09	0.1	1.28	-17.0
CE	36.33	-0.1	5.10	7.0	-1.10	-0.8	1.43	-7.2
Y, A*1:*P								
Obs	26.59		6.68		-0.10		-0.55	
PSE	26.79	0.8	6.56	-1.7	-0.17	-60.1	-0.52	5.2
CE	26.44	-0.6	6.73	0.8	-0.13	-21.6	-0.53	4.1
***Y, A*2**								
Obs	26.59		6.68		-0.10		-0.55	
PSE	26.82	0.9	6.52	-2.3	-0.18	-70.2	-0.51	7.5
CE	26.45	-0.5	6.66	-0.2	-0.15	-48.6	-0.51	8.1
***Y, A*3**								
Obs	26.59		6.68		-0.10		-0.55	
PSE	26.91	1.2	6.49	-2.8	-0.17	-68.7	-0.51	7.8
CE	26.56	-0.1	6.63	-0.7	-0.15	-42.0	-0.50	8.6

Comparisons of the first four moments. Another summary comparison of the predictions of CE and PSE concerns the predictions of the mean, standard deviation, skewness, and kurtosis of the observed frequency distributions. These moments may be viewed as four different ways of smoothing and summarizing the frequencies. The values of these moments are given in Tables 7 and 8. In addition, Tables 7 and 8 also show the *percent relative differences* (% Rel. Dif.) between the observed and predicted moments. The percent relative difference is the observed moment minus the predicted moment divided by the *absolute value* of the observed moment. Thus, positive relative differences indicate *over-prediction*, while negative values indicate *under-prediction*. In Tables 7 and 8, the relative differences have been multiplied by 100 to express them as percents. These comparisons are done separately for $\boldsymbol{X}$ or $\boldsymbol{X1}$ in $\boldsymbol{Q}$ and for $\boldsymbol{Y}$ or $\boldsymbol{Y1}$ in $\boldsymbol{P}$. Table 7 illustrates the external anchor cases and Table 8 the internal anchor cases.

Several overall tendencies are revealed by a comparison of the CE and PSE predictions of the first four moments of the observed frequencies. First, in almost every case in Tables 7 and 8, in terms of the absolute value of the percent relative difference, the CE predictions are closer to the observed data than are the PSE predictions (the few exceptions are shown in shaded cells). The predictions of the means are quite accurate for both methods; the means have the consistently smallest percent relative differences in Tables 7 and 8, but the relative differences for the CE predictions are always smaller. For the standard deviations, the percent relative differences are generally a little larger, but again, those for CE are always smaller. The percent relative differences for both sets of predictions are generally larger for the skewness and kurtosis than for the mean and variance. However, both the signs and the magnitude of the predictions for skewness and kurtosis are correct for both CE and PSE. Moreover, the few cases where PSE has the smaller percent relative difference occur for skewness and kurtosis.

As seen earlier for the goodness-of-fit measures, there is a consistent tendency for the accuracy of the predictions of PSE for the means and standard deviations to increase as the length of the anchor test increases. These tendencies are less consistent for the skewness and kurtosis predictions. The CE predictions for the mean and standard deviation for the external anchor test do not show the same consistent improvement as the length of the anchor test increases. This difference in trends for CE and PSE echoes the findings for the goodness-of-fit measures given earlier, and it is restricted to the external anchor case.

Table 8

The First Four Moments of the Observed and the Predicted Distributions and Their Relative Differences. Internal Anchor Test Cases

Tests, anchor; obs/pred.	Mean	Mean % rel. dif.	SD	SD % rel. dif.	Skewness	Skew % rel. dif.	Kurtosis	Kurt % rel. dif.
X*1, *A*1: *Q								
Obs	53.38		8.04		-0.86		0.66	
PSE	53.17	-0.4	8.47	5.3	-0.87	-1.0	0.60	-9.8
CE	53.31	-0.1	8.36	3.9	-0.86	-0.4	0.60	-8.9
***X*1, *A*2**								
Obs	53.38		8.04		-0.86		0.66	
PSE	53.11	-0.5	8.57	6.5	-0.84	1.9	0.51	-22.5
CE	53.29	-0.2	8.45	5.0	-0.83	2.8	0.51	-22.6
***X*1, *A*3**								
Obs	53.38		8.04		-0.86		0.66	
PSE	52.94	-0.8	8.67	7.8	-0.85	0.7	0.57	-14.2
CE	53.15	-0.4	8.53	6.0	-0.85	1.5	0.58	-12.0
Y*1, *A*1: *P								
Obs	42.62		10.31		-0.19		-0.56	
PSE	42.82	0.5	10.17	-1.3	-0.23	-25.3	-0.54	4.7
CE	42.62	0.0	10.27	-0.3	-0.21	-15.1	-0.54	4.6
***Y*1, *A*2**								
Obs	42.62		10.31		-0.19		-0.56	
PSE	42.89	0.6	10.18	-2.2	-0.25	-36.3	-0.51	8.8
CE	42.65	0.1	10.17	-1.3	-0.24	-29.7	-0.51	9.0
***Y*1, *A*3**								
Obs	42.62		10.31		-0.19		-0.56	
PSE	43.08	1.1	10.00	-3.0	-0.25	-35.2	-0.51	8.9
CE	42.81	0.4	10.12	-1.8	-0.24	-28.0	-0.51	9.5

Conclusions and Discussion

This study investigates the assumptions that underlie two of the most used OSE methods for the NEAT design—CE and the PSE. In the usual operational settings, these assumptions are untestable and cannot be evaluated. In this study we used a special data set that allowed us to test the predictions that the assumptions of both CE and PSE make regarding the data that are missing in real NEAT designs.

We found that the two methods were very similar in terms of how well the predicted distributions approximated the observed distributions, with the CE-based results being slightly closer to the observed distributions than those of PSE. In addition, we observed that while the predictions of PSE were consistently improved using the longer and more reliable anchor tests, this was not found consistently for CE. The lack of the expected trend for CE appears for the goodness-of-fit measures (Table 6) and for the means and SDs (Table 7), but only for the external anchor case. In reviewing an earlier draft of this paper, our colleague, Tim Moses, did additional analyses with these data and found that when the PSE and CE predicted frequencies were compared to *smoothed* versions of the observed frequencies obtained by fitting loglinear models to them, the trends for CE in Table 6 became consistent with those of PSE. These results suggest that the lack of this trend for CE may be due to sampling variability.

In retrospect, we recognize that we could have approached the problem of putting the predictions of CE and PSE on the same footing in a different way. Instead of *discretizing* the CE-based continuous cdfs for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$, we could have *continuized* the smooth, PSE-based, predicted frequencies for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$. These results could then be compared with *criterion cdfs* based on the observed frequencies for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$. Finding the criterion cdfs for $\boldsymbol{X}$ and $\boldsymbol{Y}$ would have required the addition of *presmoothing* the observed frequencies for $\boldsymbol{X}$ in $\boldsymbol{Q}$ and $\boldsymbol{Y}$ in $\boldsymbol{P}$ and then *continuizing* them to get the criterion cdfs. With the predicted cdfs for CE and PSE and the criterion cdfs in hand, we could then have compared these cdfs in various ways. Such an approach is interesting and worth further consideration, but it would involve more presmoothing and continuizing than the approach we took here.

In discussing this study, we wish to mention two further issues that have arisen.

Discretizing continuous distributions. The method for discretizing the continuous cdfs of CE given in (12) and (13) has uses beyond obtaining predicted discrete score distributions for CE. In von Davier, et al. (2004a), it is proposed to use the *percent relative error* (PRE) in the

moments of $\boldsymbol{Y}$ and the transformed $\boldsymbol{X}$-scores, $e_{\boldsymbol{Y}}(\boldsymbol{X})$ on the target population, $\boldsymbol{T}$, as a way of diagnosing the adequacy of an equipercentile equating function. The PRE measures they use have the form

$$\text{PRE}(p) = 100[\mu_p(\text{e}_{\boldsymbol{Y}}(\boldsymbol{X})) - \mu_p(\boldsymbol{Y})]/\mu_p(\boldsymbol{Y}) \tag{19}$$

where

$$\mu_p(\boldsymbol{Y}) = \sum_k (y_k)^p s_{k\boldsymbol{T}} \text{ and } \mu_p(\text{e}_{\boldsymbol{Y}}(\boldsymbol{X})) = \sum_j e_{\boldsymbol{Y}}(x_j)^p r_{j\boldsymbol{T}} \ . \tag{20}$$

In (20), $s_{k\boldsymbol{T}}$ denotes the discrete score probabilities for $\boldsymbol{Y}$ on $\boldsymbol{T}$, while $r_{j\boldsymbol{T}}$ denotes them for $\boldsymbol{X}$ on $\boldsymbol{T}$. $\mu_p(\boldsymbol{Y})$ and $\mu_p(\text{e}_{\boldsymbol{Y}}(\boldsymbol{X}))$ are the p^{th} moments of $\boldsymbol{Y}$ and the transformed $\boldsymbol{X}$ scores, $\text{e}_{\boldsymbol{Y}}(\boldsymbol{X})$, over the target population. The values of p run from 1 to 10 in von Davier, et al. (2004a) These authors pointed out that, in trying to apply PRE(p) to the CE method, they were hampered by the fact that CE does not estimate the discrete score distributions $r_{j\boldsymbol{T}}$ and $s_{k\boldsymbol{T}}$. However, PSE does produce such estimates.

What CE *does* produce are the *continuized* cdfs of $\boldsymbol{X}$ and of $\boldsymbol{Y}$ on $\boldsymbol{T}$, in a manner similar to equations (11) and (12). The discretizing method in (13) and (14) worked so well to produce the accurate predictions of CE in this study that we think that it may be a useful way to produce the discrete probabilities needed for computing PRE(p) values for CE. The discretization would be applied to the CE values for $F_{\boldsymbol{T}}(x)$ and $G_{\boldsymbol{T}}(y)$ in that case. We believe that this is a useful area for future research.

As a final comment about the discretizing method in (13) and (14), we note that it can be shown to have the following reciprocal property with the linear-interpolation method of continuizing discrete distributions that is traditionally used for equipercentile equating (Kolen & Brennan, 2004). If a discrete distribution is continuized by linear interpolation and then discretized using (13), the original discrete distribution is the result. This finding was verified by our colleague Tim Moses who did the calculation directly on the data in this study. If the Gaussian kernel smoothing method is used to continuize the distributions and the bandwidths are chosen to make the densities of the continuous cdfs close to the presmoothed frequencies, then the discretizing method in (13) and (14) will very closely reproduce the original presmoothed frequencies, but there will be tiny differences, as Moses found.

Pseudo-test studies versus simulation studies. After all the work that was done to create pseudo-tests and pseudo-anchor tests, we wonder if it was worth it. A great deal of care was taken to make sure that the pseudo-tests, ***X*** and ***Y***, were parallel in content covered, but sufficiently different in difficulty to require equating. The result was two tests that had quite different mean scores on all of the populations. In an IRT-based simulation study there is no issue of content coverage (at least for a one-dimensional latent variable) and a difference in test difficulty is just a matter of the choice of difficulty parameters. Several different differences in test difficulty could be easily studied in a simulation study. To do this in a pseudo-test study would require attention to the overlap of the test items in the pseudo-tests that would force correlations between the results for different pairs of ***X***s and ***Y***s. In addition, in the pseudo-test study, considerable effort was made to select appropriate test items for the anchor tests, ***A*1**, ***A*2,** and ***A*3**. They were all chosen to be representative of the original test's content and difficulty. Furthermore, the variation in the length of the anchors was intended to vary their reliabilities. What resulted was a very modest range of reliability differences. In a simulation study, a wider range of anchor test reliability could have been achieved in a variety of ways. The representativeness of the content of the anchor tests would not arise in a simulation study.

The pseudo-test study was designed to make CE and PSE produce *different* results. ***P*** and ***Q*** had mean scores on the anchor tests that were different enough to be a cause for concern in an actual equating. This difference is known to make CE and PSE differ. The disappointment is that, in retrospect, this effort made for only little differences between CE and PSE. Both methods make very similar predictions and are more similar to each other than they are to the data. It is true that CE performed a bit better than PSE did, but the striking finding is how little difference there is between the two methods in this study. A simulation study would be easier to mount and more differences could be arranged to produce cases where CE and PSE are more different. We believe that larger differences between CE and PSE would lead to sharper tests of the untestable assumptions that underlie these methods.

In criticizing pseudo-test studies, we recognize that there are also clear benefits to them. The most notable is that the data are real rather than made up. The item responses reflect the behavior of real examinees rather than a statistical model. This is an important benefit, but in the present case we wonder if it is important enough to overcome the drawback of the labor-intensive construction of pseudo-tests and the lack of control this leads to for important factors

that could be more easily varied in a simulation study. Careful design of simulation studies can mimic many features of real data from examinees, and simulations can serve as a type of “animal model” for real test data.

References

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic Press.

von Davier, A. A. (2003). *Notes on linear equating methods for the Non-Equivalent Groups design* (ETS RR-03-24). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus poststratification methods for equating and test linking. In N. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program® Examinations* (ETS RR-03-27). Princeton, NJ: ETS.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The chain and poststratification methods for observed-score equating: their relationship to population invariance. In N. J. Dorans (Ed.), Assessing the population sensitivity of equating functions [Special issue] *Journal of Educational Measurement*, *41*, 15-32.

von Davier, A. A., & Ricker, K. (2006). *The role of the anchor test in a non-equivalent group design.* Paper to be presented at AERA 2006, San Francisco.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006), *An evaluation of the kernel equating method. A special* study with pseudo-tests constructed from real test data (ETS RR-06-02). Princeton, NJ: ETS.

Dorans, N.J., Liu, J., & Hammond, S, (2005, April). *The role of the anchor test in achieving population invariance across subpopulations and test administrations.* Paper presented at the annual meeting of the NCME, San Diego, CA.

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.

ETS. (2005). KE-Software [Computer software]. Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

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.

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 NCME, San Francisco.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS RR-93-04). Princeton, NJ: ETS.

Appendix

Graphs of Observed and Predicted Frequencies and Their Freeman-Tukey Residuals for the Other Cases Examined in the Study

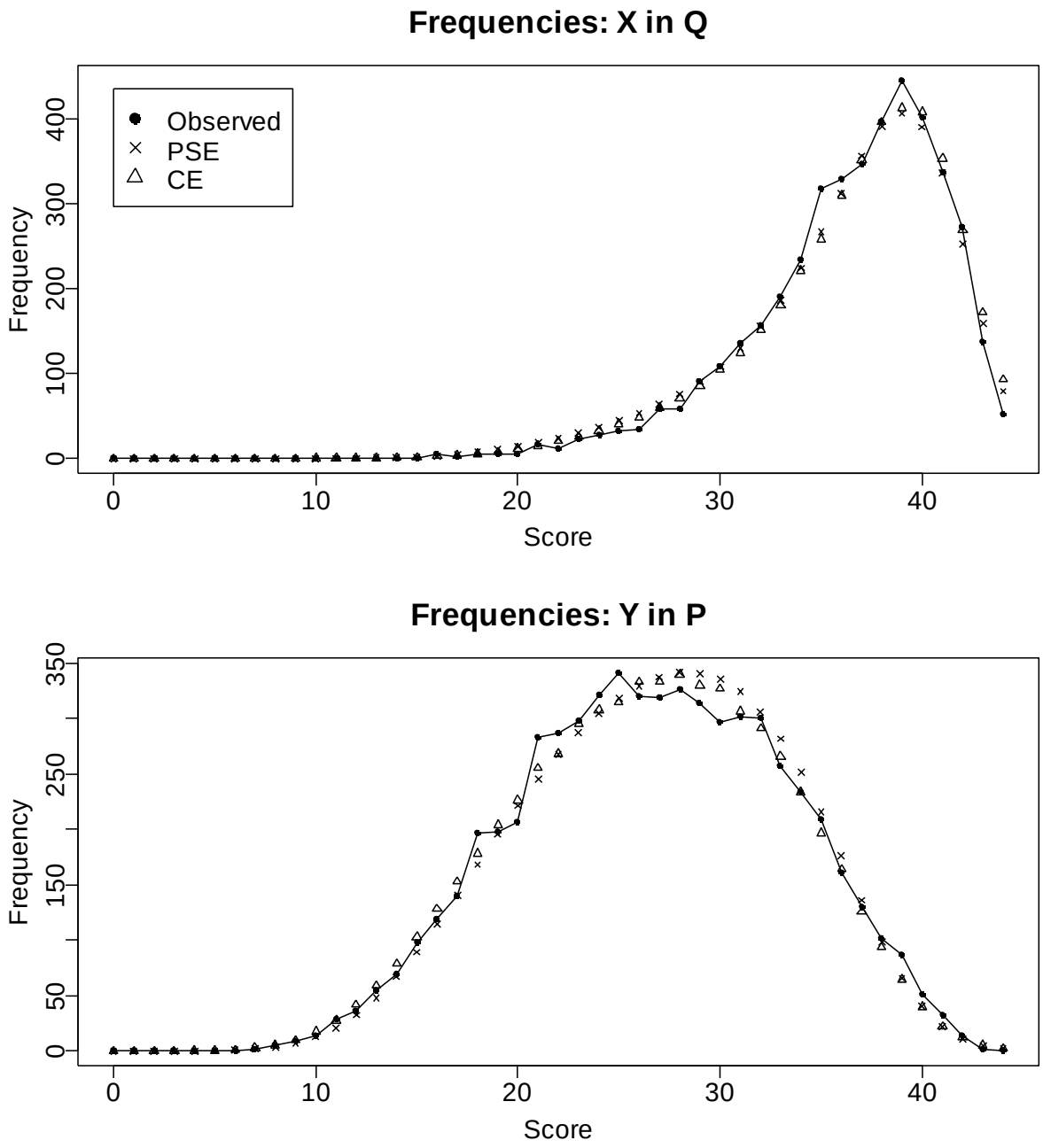

Figure A1. **Frequencies for *X* in *Q* and *Y* in *P* for external anchor test *A*2.**

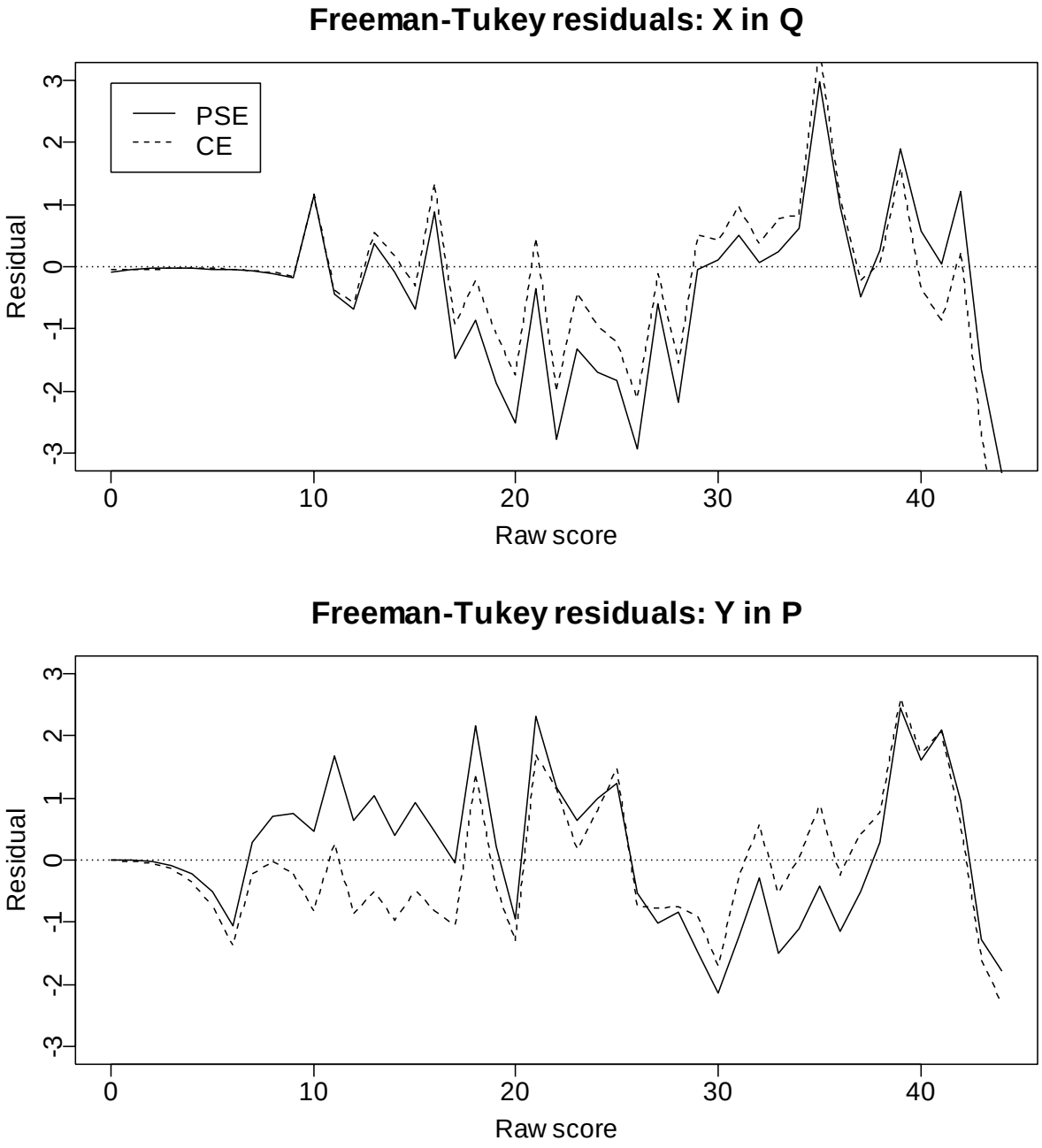

Figure A2. **Freeman-Tukey residuals for *X* in *Q* and *Y* in *P* for external anchor test *A*2.**

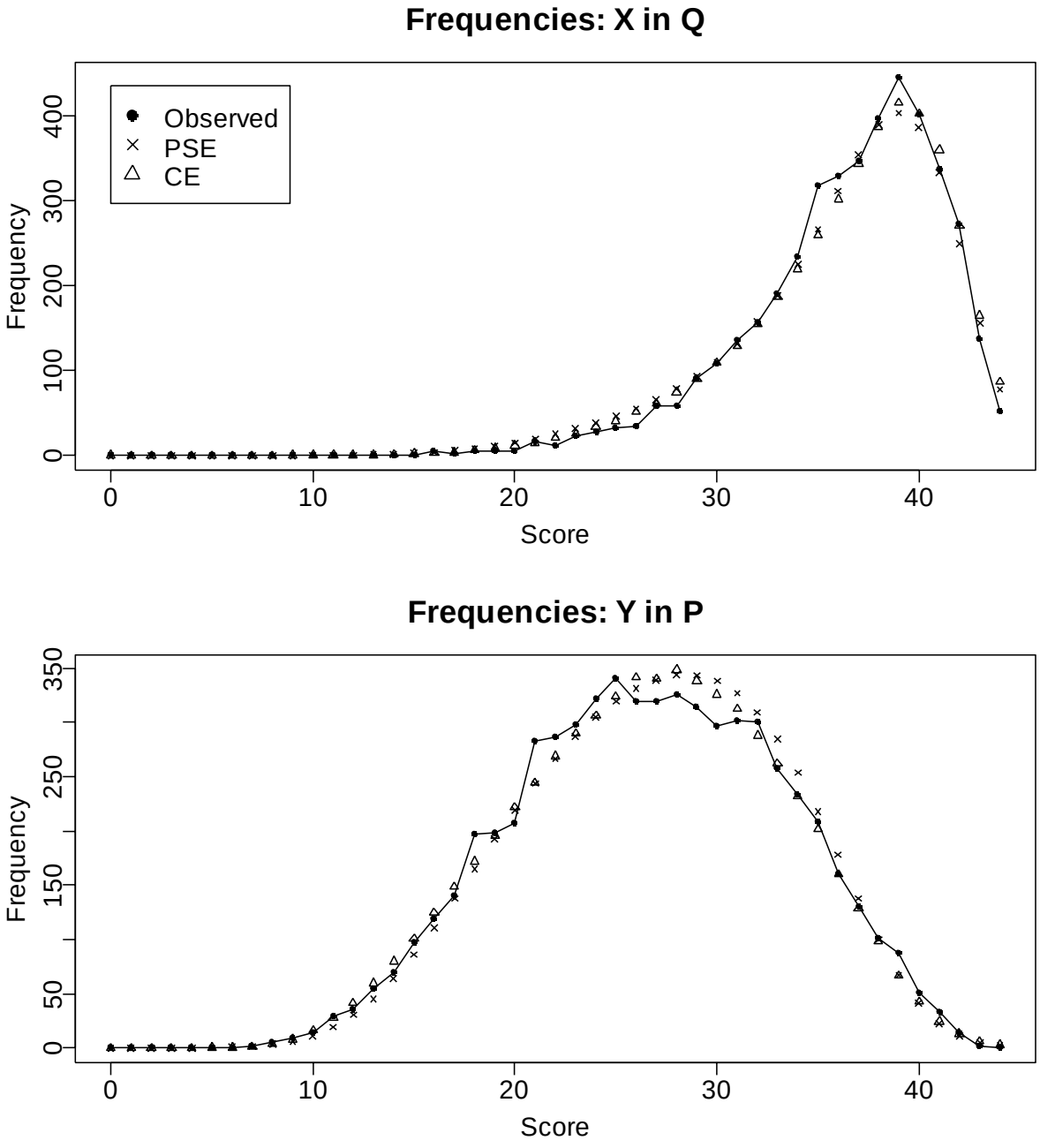

Figure A3. **Frequencies for *X* in *Q* and *Y* in *P* for external anchor test *A*3.**

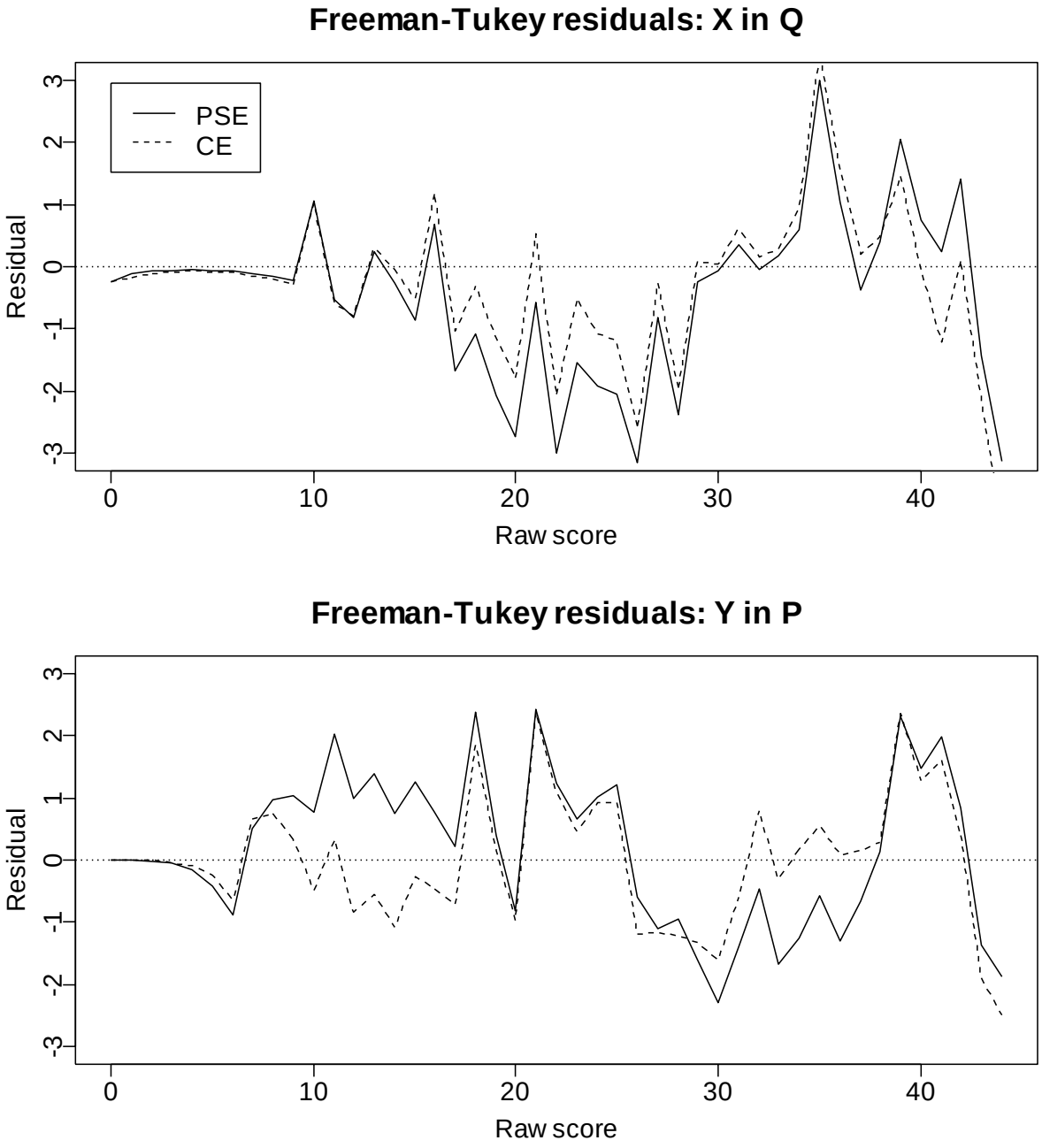

Figure A4. **Freeman-Tukey residuals for *X* in *Q* and *Y* in *P* for external anchor test *A*3.**

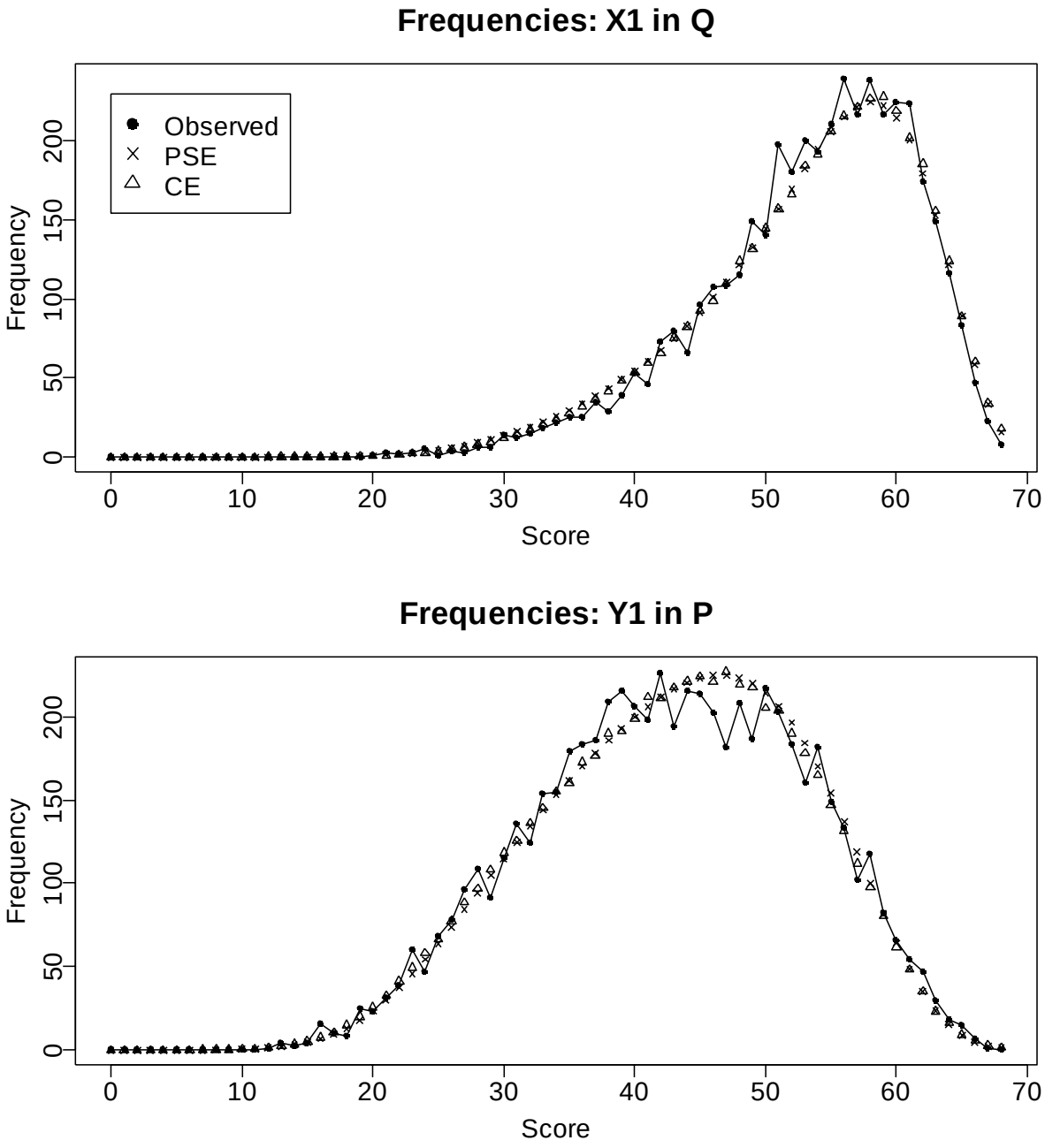

Figure A5. **Frequencies for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*2.**

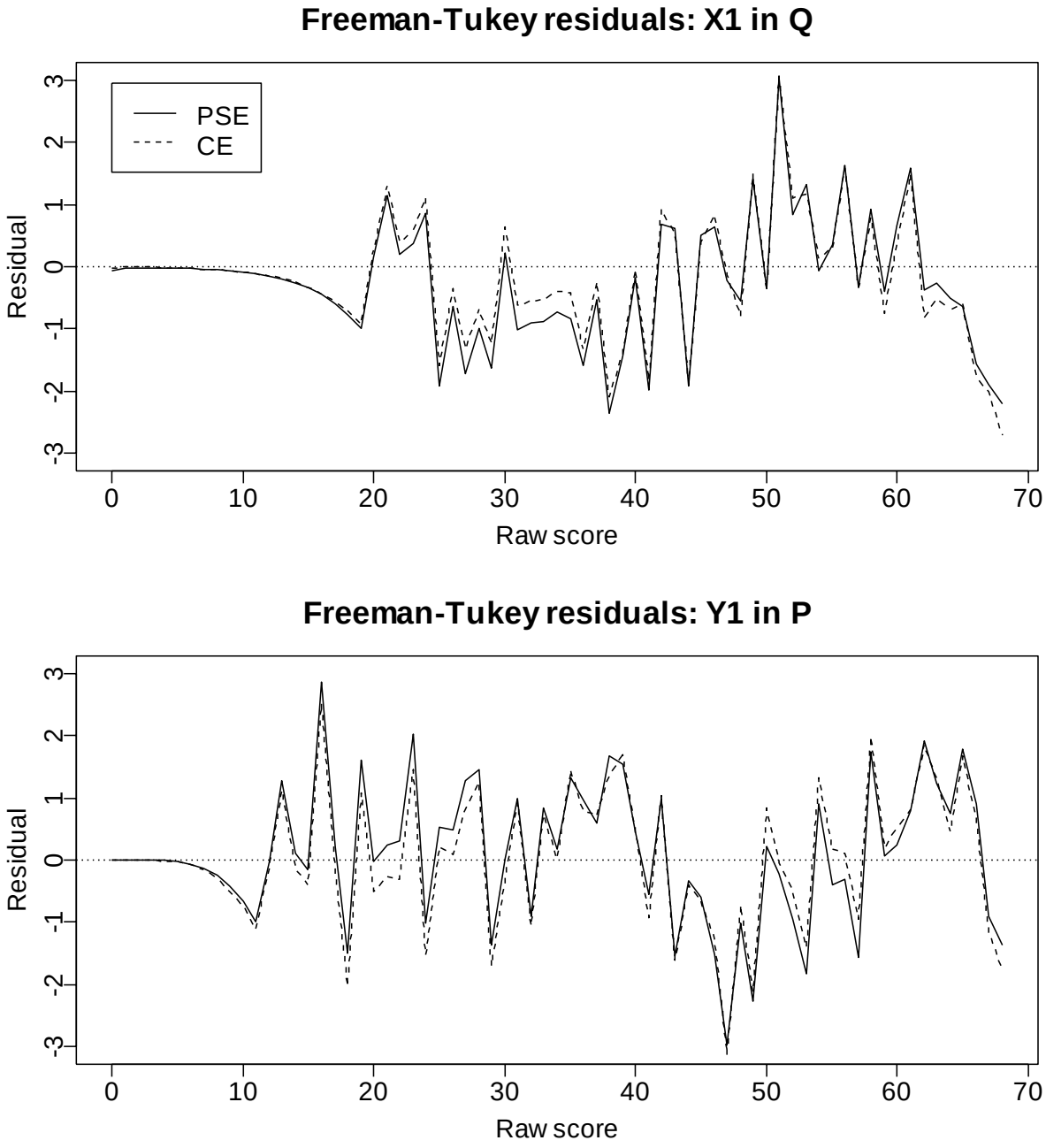

Figure A6. **Freeman-Tukey residuals for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*2.**

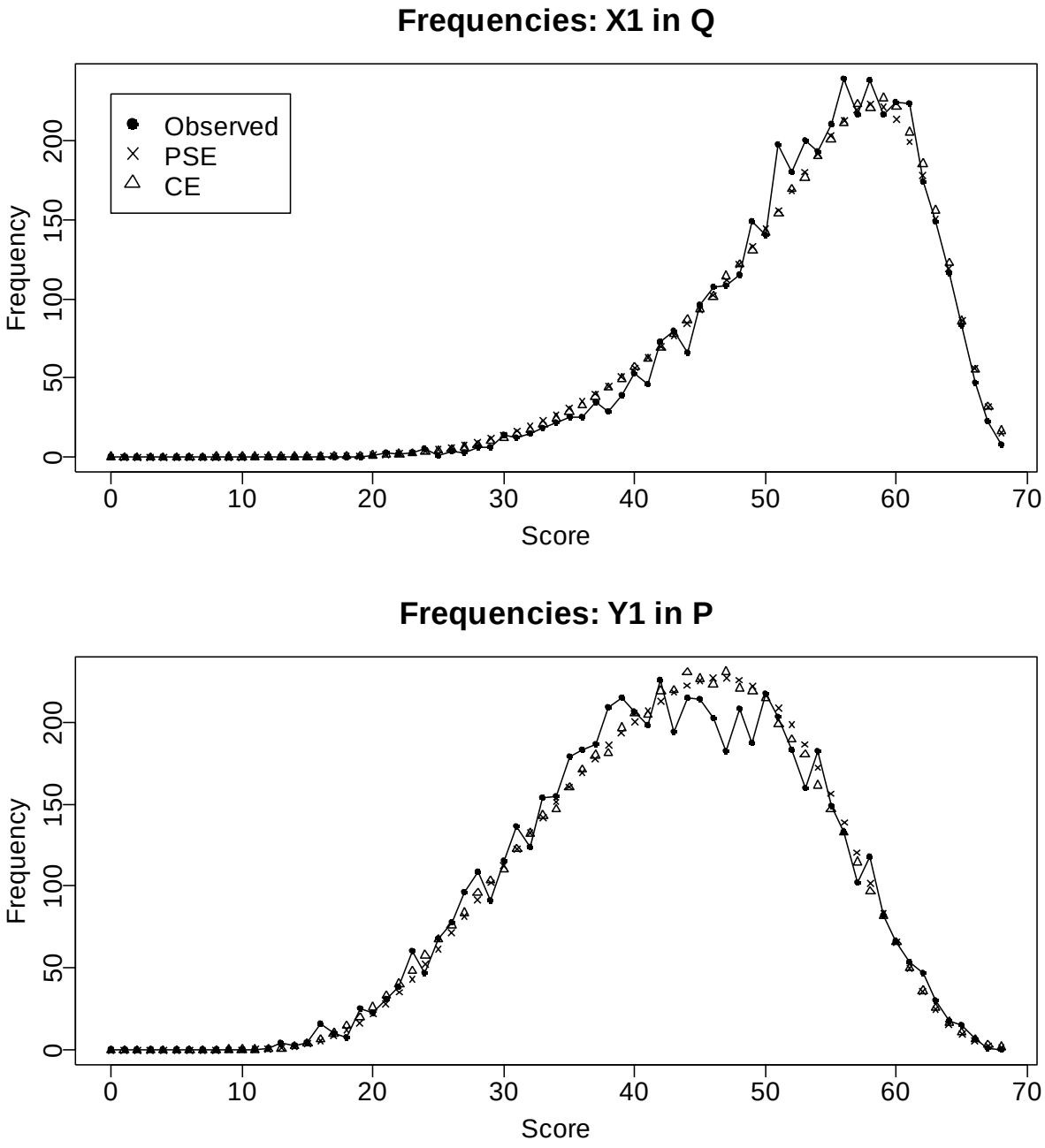

Figure A7. **Frequencies for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*3.**

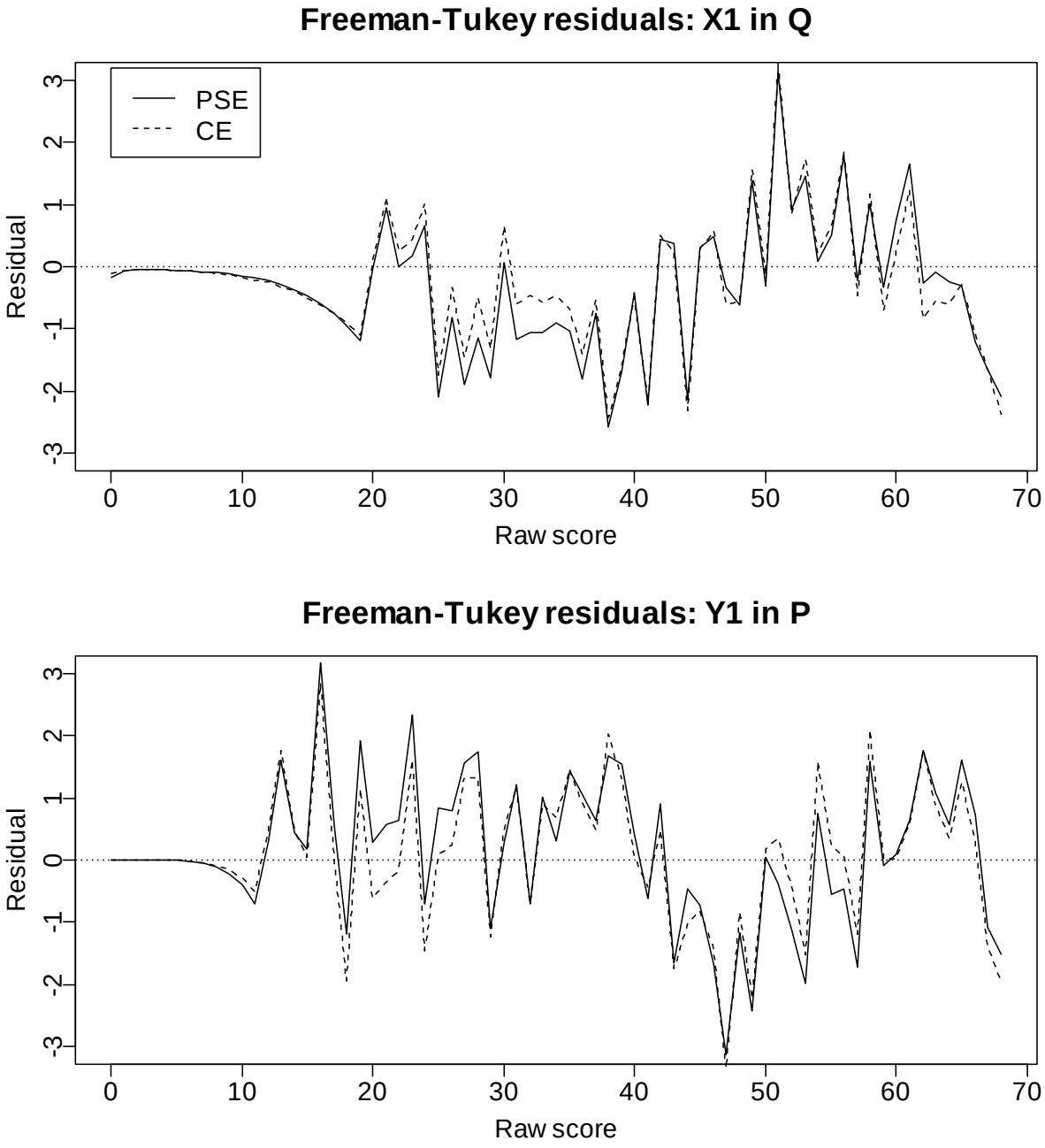

Figure A8. **Freeman-Tukey residuals for *X*1 in *Q* and *Y*1 in *P* for internal anchor test *A*3.**