Research Report

# An Evaluation of the Kernel Equating Method: A Special Study With Pseudotests Constructed From Real Test Data

Alina A. von Davier

Paul W. Holland

Samuel A. Livingston

Jodi Casabianca

Mary C. Grant

Kathleen Martin

Research & Development

March 2006
RR-06-02

# An Evaluation of the Kernel Equating Method:
# A Special Study With Pseudotests Constructed From Real Test Data

Alina A. von Davier, Paul W. Holland, Samuel A. Livingston,
Jodi Casabianca, Mary C. Grant, and Kathleen Martin
ETS, Princeton, NJ

March 2006

## Abstract

This study examines how closely the kernel equating (KE) method (von Davier, Holland, & Thayer, 2004a) approximates the results of other observed-score equating methods—equipercentile and linear equatings. The study used pseudotests constructed of item responses from a real test to simulate three equating designs: an equivalent groups (EG) design and two non-equivalent groups with anchor test (NEAT) designs, one with an internal anchor and another with an external anchor. To compare results, the study sets the equating function in the EG design as the equating criterion. In these examples, the KE results were very close to the results from the other equating methods. Moreover, in almost all situations investigated, the KE results were closer to the equating criterion.

Key words: Kernel equating, observed-score equating methods, non-equivalent groups with anchor test (NEAT) design, the equivalent groups (EG) design

## Acknowledgments

The authors thank Tim Moses, Ning Han, Stacie Rupp, and Stephanie Fournier-Zajac for their help with additional computations. The authors also thank Dan Eignor, Neil Dorans, Michael Walker, and Michael Kolen for their suggestions and comments on earlier versions of the report, and Kim Fryer for editorial help.

## Table of Contents

## List of Tables

# Introduction

Kernel equating (KE; Holland & Thayer, 1989; von Davier, Holland, & Thayer, 2004) is an equipercentile observed-score equating procedure in which the score distributions to be equated are converted from discrete distributions to continuous distributions by using a normal (Gaussian) *kernel*, as opposed to using linear interpolation as in the traditional equipercentile equating method. See Appendix A for a brief description of the KE method.

KE holds the promise of approximating several commonly used observed-score equating methods while providing for the routine calculation of new and previously unavailable measures of statistical accuracy. KE can achieve the former because it is based on a flexible family of equipercentile-like equating functions that contains the linear equating function as a special case.

This report describes a research study intended to answer two questions about KE: (a) Are the results of KE at least as close to an equating criterion as those of other methods? and (b) How closely do the results of KE approximate the results of other equating methods?

The data for this study are item responses from real examinees taking a real test. The criterion equating against which we evaluate the accuracy of all the other equatings is an equivalent-groups (EG) classical equipercentile equating that makes use of all the examinees responses (the combined group). The equatings evaluated are equating functions from the EG design (which is based on the combined group and described in detail later in this report) and from two nonequivalent groups with anchor test (NEAT) data collection designs, one with an external anchor and one with an internal anchor. The equatings in the NEAT designs were computed from partial data sets, structured to create a situation in which two groups of examinees who are unequal in ability take different forms of a test that are unequal in difficulty. Hence, the goal of the study was to determine the degree to which KE can approximate the results of other observed-score equating methods (equipercentile and linear for the EG design and the Tucker, Levine observed-score, chained equipercentile, chained linear, and frequency estimation equating methods for the two NEAT designs).

The body of the report focuses on the results for the NEAT design with an external anchor test. The results for the internal anchor are similar to those for the external anchor and are briefly reported in Appendix C.

### *A Previous Evaluation of Kernel Equating*

Some features of the present study have been adapted from a previous study that evaluated kernel equating (Livingston, 1993). In that study, pairs of equating samples of a specified size were randomly drawn from a group of more than 93,000 examinees who had taken a 100-item test. From the 100 items on that test, the investigator constructed two overlapping 58-item test forms, similar in content but differing in difficulty. In order to simulate a NEAT design, the scores on one of the two forms was treated as unknown or missing, as if each examinee had a score on only one of the forms and an anchor test score. The criterion equating was the equipercentile equating of the two forms, using the responses of all examinees. For the KE and the chained equipercentile functions, loglinear models were fitted to the discrete score distributions prior to equating (i.e., the distributions were presmoothed; see Holland & Thayer, 1989, 2000). The study compared the accuracy of five equating functions: three kernel equatings (using different bandwidth parameters), a chained equipercentile equating of the smoothed discrete score distributions, and a chained equipercentile equating of the observed (unsmoothed) score distributions. In Livingston's study, the equated scores produced by kernel equating were slightly more accurate than those produced by the chained equipercentile equating of the smoothed discrete distributions and much more accurate than those produced by the chained equipercentile equating of the observed (unsmoothed) distributions. The kernel equating results differed only slightly across values of the bandwidth parameter, except near the ends of the score range, where the large bandwidth value produced biased results.

The rest of this report is structured as follows: The next section introduces the notation and discusses the equating methods investigated; next, the report describes the equatings in the EG design and the NEAT design with an external anchor; and then the report compares the equating results across different functions. The last section presents the conclusions drawn by the study investigators.

## Notation

There are two test forms to be equated, $X$ and $Y$, and a target population, $T$, for which the scores on the two test forms are to be made equivalent (for the population as a whole, not necessarily for every individual in the population). In the EG design, the two operational tests to be equated, $X$ and $Y$, are given to two samples of examinees from the same population, $T$. In the NEAT design, $X$ and $Y$ are given to two samples of examinees from different test populations or administrations (denoted here by the populations $P$ and $Q$). In addition, an anchor test, $A$, is given to both samples from $P$ and $Q$.

The target population, $T$, for the NEAT design is assumed here to be a *mixture* of $P$ and $Q$ in which $P$ and $Q$ are regarded as partitioning $T$. $P$ and $Q$ are given weights that sum to 1. This is denoted by

$$T = wP + (1 - w)Q. \tag{1}$$

See Kolen and Brennan (2004) and Livingston (2004) for a detailed discussion of the concept of the target population.

It is always a good idea to be explicit as to what $T$ is in an equating design. For example, in this study, where we are interested in evaluating the KE method in a NEAT design, we consider the target population to be as in (1). This implies that the criterion equating (i.e., the equating function to which we compare all the equating results) should be on the same population. Therefore, the criterion equating design, the EG design in this case, was computed by pooling the data from the two administrations, insuring that the target population is of the form (1), with the $w$ determined by the relative size of the samples from $P$ and $Q$ (i.e., $w = nP/(nP + nQ)$), where $nP$ and $nQ$ are the sample sizes of the samples from $P$ and $Q$, respectively). The score distributions computed for $P$ and $Q$ separately are weighted by $w$ and $(1 - w)$ to obtain distributions of these same quantities for $T$. Two reviewers suggested that in order to define an equating criterion we should check if the criterion equating in the combined group is the same as the equatings inside each of the groups. While we provided the results of these additional equatings (see Appendix B), we consider that these analyses check a population invariance assumption and cannot influence the choice of the criterion. The choice of the equating criterion is based on a decision about the appropriate target population as described in (1) and the appropriate shape of the equating function.

Many observed-score equating methods are based on the equipercentile equating function. It is defined on the target population, $T$, as:

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

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

Linear equating assumes that $F_T(x)$ and $G_T(y)$ are continuous and have the same shape while possibly differing in mean and variance. The linear equating function, $\mathrm{Lin}_{Y;T}(x)$, is defined by

$$\mathrm{Lin}_{Y;T}(x) = \mu_{YT} + \sigma_{YT}((x - \mu_{XT})/\sigma_{XT}). \tag{3}$$

In Theorem 1 of von Davier, Holland, and Thayer (2004a, 2004b), it is shown that any equipercentile equating function can be decomposed into the corresponding linear equating function and a nonlinear part. The corresponding linear function is a function that relies on the same assumptions as the equipercentile function—for example, chained linear and chained equipercentile functions rely on the same assumptions.

## Equating Methods

### *EG Design*

In an EG design, there are two categories of equating functions, equipercentile and linear. KE can approximate both functions by manipulating the bandwidth: KE with optimal bandwidths approximates the equipercentile equating function, and KE with large bandwidths approximates the linear equating (see Appendix A). Table 1 lists the equating methods computed in the EG design. The left column lists the equating methods that are not kernel equatings. The right column indicates, for each of these methods, the version of kernel equating that was expected to produce a close approximation to it.

**Table 1**

***Equivalent Groups Design Equating Methods and Corresponding KE Procedures***

| Equating method | KE version |
|---|---|
| Classical equipercentile | KE, with optimal bandwidths |
| Classical linear | KE, with large bandwidths |

### *NEAT Design*

In an anchor equating design, there are three fundamentally different ways to use the information provided by the anchor in an observed-score equating setting. The anchor score can be used as a conditioning variable (i.e., a covariate) for estimating the score distributions or statistics on the tests to be equated. This approach is similar to poststratification in survey research, and we will refer to equating methods based on this approach as poststratification equating (PSE). A second way to use the anchor information is to have the anchor score as the middle link in a chain of symmetric linking relationships. We will refer to equating methods based on this approach as chained equating (CE). The equating functions in these two categories can be linear or nonlinear in shape. The third way to use the anchor information leads to another method available in the NEAT design—the Levine observed-score linear equating method, which is based on estimated relationships between true scores on the test forms to be equated and on the anchor. Table 2 illustrates these three classes of equating methods for the NEAT design.

**Table 2**

***Observed-Score Equating Functions for the NEAT Design***

| | PSE | CE | Levine |
|---|---|---|---|
| Linear | Tucker<br>Braun and Holland | CE linear | Levine observed-score equating |
| Nonlinear | Frequency estimation | CE equipercentile | None |

*Poststratification equating methods.* The linear poststratification methods used in this study include the Tucker method and the Braun and Holland method (see Braun & Holland, 1982; Kolen & Brennan, 2004).

The PSE linear or Braun and Holland linear method uses the poststratified score probabilities to compute the mean and variance of $X$ and $Y$ on $T$. These moments are then used to directly compute $\mathrm{Lin}_{Y;T}(x)$, defined in (3). This approach makes different assumptions than does the Tucker linear method, though they are related. In Tucker linear equating, the assumptions regarding population invariance are weaker because they refer only to the first two moments of the conditional distributions of $X$ and $Y$ given $A$. However, the additional assumptions of the Tucker method (i.e., linear regression and constant conditional variance) are stronger in the sense that linear PSE can have nonlinear regressions of $X$ and $Y$ on $A$ and nonconstant conditional variances as well. The form of these conditional moments depends on the form of the model used to presmooth the bivariate data tables that arise in the NEAT design (see Holland &Thayer, 2000). Hence, the Tucker linear method and the PSE linear will agree well only in specific circumstances, when the regressions of $X$ and $Y$ given $A$ are approximately linear and the conditional variances are almost constant.

A poststratification method that allows for a curvilinear equating relationship is frequency estimation equipercentile equating.

The kernel version of poststratification equating provides approximations to frequency estimation equipercentile and to the Braun and Holland linear methods (see von Davier et al., 2004b, for the theoretical proof). When optimal bandwidths are chosen to closely approximate the discrete distribution, then the kernel version of poststratification equating will approximate a frequency estimation equipercentile equating computed from the presmoothed bivariate distributions (the presmoothing is accomplished using loglinear models—see Holland & Thayer, 2000). When large bandwidths are chosen, then the result will approximate the Braun and Holland linear method.

*Chained equating methods.* The chained equating methods used in this study are the chained linear method and the chained equipercentile method. The formulas for these methods (except for the Braun-Holland method) are presented concisely in Angoff (1984) and explained, without formulas, in Livingston (2004). For a complete presentation, see Kolen and Brennan (2004). The chained equating represents a chain of linking from $X$ to $A$ and from $A$ to $Y$. In general, if each of the two links is equipercentile, then the final equating is equipercentile as well. There are some other cases, but they are beyond the scope of this study. If each of the two links is linear, then the final equating is linear as well. The kernel version of chained equating will approximate the chained equipercentile method when the optimal bandwidths are used and will approximate the chained linear method when large bandwidths are used.

*Levine observed-score equating method.* The Levine method does not yet have a curvilinear analogue, and there is no version of KE that approximates the Levine method. Nevertheless, we included the Levine observed-score equating method for comparison purposes, because under some circumstances it is more accurate than other linear equating methods (see Petersen, Marco, & Stewart, 1982).

Table 3 lists the anchor equating methods included in this study. The left column lists the equating methods that are *not* kernel equatings. The right column indicates, for each of these methods, the version of kernel equating that was expected to produce a close approximation to it. As mentioned previously, the KE PSE approximates the Braun and Holland linear method, not the

Tucker method. The former two methods agree in most cases when the regressions of *X* and *Y* given *A* are approximately linear and the conditional variances are almost constant.

**Table 3**

***Anchor Equating Methods and Corresponding KE Procedures***

| Equating method | KE version |
|---|---|
| Chained equipercentile | KE chained, with optimal bandwidths |
| Frequency estimation equipercentile | KE poststratification, with optimal bandwidths |
| Chained linear | KE chained, with large bandwidths |
| Tucker | Not directly available (KE poststratification with large bandwidths under certain conditions) |
| Braun and Holland linear (not available) | KE poststratification with large bandwidths |
| Levine observed-score | (None) |

In this study, there are certain limitations to comparing the equating results due to the limitations of the software available—that is, GENASYS (ETS, 2004). We also used the newly created KE software (ETS, 2004) for computing the KE chained linear and equipercentile functions and a SAS macro (Moses, von Davier, & Casabianca, 2004) for the loglinear presmoothing in order to obtain the appropriate input for the KE software. The Braun and Holland linear method outside the KE framework is not available.

The description of the equating results will include the continuization values of the KE functions (see Appendix A, Step 4), the diagnostic measures, the percentage of relative errors, which are available only in the KE framework (see also Appendix A, Step 4), and the standard errors of equating (SEE).

## Procedure for the Study

The evaluation of any equating method requires an equating design, where the equating criterion in the target population is known (Harris & Crouse, 1993). In practice, it is very difficult to find a known criterion for equating. In this study, we used real responses from real people taking real tests as raw material to construct the tests, the equating design, and the criterion equating for the study.

### *Data*

The data for this study came from one form of a licensing test for prospective teachers of children in elementary school. The test 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. The mean total scores of the examinees taking the test at these two administrations, *P* and *Q*, differed by approximately one fourth of a standard deviation, as can be seen in Table 4.

### *Test Construction*

We used the items in the 119-item test form to create two smaller forms (each with 44 items, 11 from each of the four content areas) parallel in content but differing in difficulty to be given with a representative set of 24 items in common (6 from each content area) to provide an (external) anchor for equating.

The anchor test is treated as an external anchor in the body of the report and as an internal anchor in Appendix C.

**Table 4**

***Comparison of the Examinees at the Two Administrations for the Initial Test***

| Administration | *P* | *Q* |
|---|---|---|
| Number of examinees | 6,168 | 4,235 |
| Mean | 82.33 | 86.16 |
| SD | 16.04 | 14.19 |

The anchor set was constructed to cover the content tested by the three tests, the 119-item test, and the two 44-items tests (to the extent possible) and to represent the content categories in the same proportions as in a full-length test. The anchor contains a set of items that has a mean difficulty approximately equal to the recommended mean for the long test and a difficulty range that is also approximately the same as that for the whole long test.

The reliabilities of the new tests were about 0.8; the correlations of the tests with the external anchor were 0.78 in *P* and 0.76 in *Q*. See also Appendix D for details and for the internal anchor case.

The two 44-item test forms will be referred to in this report as Form *X* and Form *Y*. Form *X* will be considered the new form, and Form *Y* the old form in all the equatings. Form *Y* was the more difficult of the two forms; the difference between the mean scores of the combined group on Forms *X* and *Y* was approximately 1.4 standard deviations. Table 5 shows the means and standard deviations of the scores on Forms *X* and *Y* for the examinees in *Q*, the examinees in *P*, and the combined group.

To summarize: Because everybody in *P* and in *Q* answered the 119 items, everybody in *P* and *Q* also answered the two 44-item tests, Form *X* and Form *Y*, that come from the 119-item tests.

We then used the samples from each of the groups of examinees as equating samples to equate the two smaller forms in an anchor equating, as if each of the smaller forms had been given at only one of the administrations.

**Table 5**

***Comparison of the Examinees at the Two Administrations for Both Forms***

| | | Examinees | | |
|---|---|---|---|---|
| | | *P* | *Q* | Combined group |
| *n* | | 4,237 | 6,168 | 10,405 |
| Mean | Form *X* | 36.4 | 35.1 | 35.6 |
| | Form *Y* | 28.0 | 26.6 | 27.2 |
| SD | Form *X* | 4.8 | 5.7 | 5.4 |
| | Form *Y* | 6.3 | 6.7 | 6.6 |

### *Equatings*

As explained earlier, the criterion equating design is an EG design obtained by pooling data from *P* and *Q* (the combined group). To provide a criterion for the correctness of the anchor equatings, we used the equipercentile equating method to equate the presmoothed distributions of scores on the two smaller forms in the EG design (i.e., the combined group of examinees from the two test administrations). We fit loglinear models to the discrete score probability distributions to presmooth the data (Holland & Thayer, 1989, 2000). We explained in a previous section how we chose the criterion for the equating methods by defining carefully the target population as being the same for both the NEAT design and the criterion-design, as in (1). We computed the equatings in the EG design by the methods described in Table 1. We computed the equatings in the NEAT design by the anchor equating methods mentioned in Table 2.

This research answers two questions posed earlier about KE: (a) Are the results of KE at least as close to an equating criterion as those of other methods? (b) How closely do the results of KE approximate the results of other equating methods? In order to answer the first question, we compared the results of all of the anchor equatings, both traditional and KE versions, with the results of the equating criterion (i.e., the equipercentile equating in the EG design). In order to answer the second question, we compared the results of the traditional equating methods and the KE versions of them both in the EG and NEAT designs.

## The Combined Group Equatings: Description

The criterion equating is an EG equipercentile equating of smoothed distributions of each of the 44-item tests (Forms $X$ and $Y$) as in (2) in the combined group (i.e., on a target population that is a weighted average of the two groups from $P$ and $Q$). The smoothing technique, loglinear smoothing, allows the user to specify how many moments of the distribution are to be preserved. After investigating the fit of the smoothed distributions based on preserving three, four, and five moments of the observed distribution, we chose the loglinear model that preserved five moments for each of the two univariate distributions, of $X$ and $Y$, respectively. The fit statistics that provided the basis for this decision included the likelihood ratio chi-square, Pearson chi-square, Freeman-Tukey residuals, Akaike information criterion, and the consistent Akaike information criterion (Bozdogan, 1987).

The optimal continuization values for the KE are $h(X) = 0.5592$ and $h(Y) = 0.6298$. For the KE linear, the large bandwidths were $h(X) = 53.70$ and $h(Y) = 65.60$.

### *Diagnosis of the KE Functions*

It is important that an equating function, as the function of the discrete $X$, matches the discrete target-distribution, $Y$. In order to assess this match, we compare up to the $10^{th}$ moments of the two distributions, $e_Y(X)$ and $Y$, via the percent-relative error in the $p^{th}$ moment (PRE) formula (see Appendix A, Step 4). The results are given in Table 6. We have such diagnostic measures *only* for the KE functions and have calculated the PRE values for both KE functions with optimal bandwidths and with large bandwidths.

The PRE values for the KE with optimal bandwidths indicate a good match between the equated function computed at the discrete values of *X* and the targeted distribution of *Y* (see Table 6). There is up to 2.9% disagreement in the 10$^{th}$ moment. The PRE values for the KE with large bandwidths show that the linear equating function at the discrete values of *X* and the *Y* matches fewer moments than the KE with optimal bandwidths. With the third moments, the disagreement is already about 1%.

**Standard Errors of Equating**

The SEEs for the KE equipercentile (with optimal bandwidths) range from 1.49 (at Score 0) to 0.10 (for Scores 40 and 41) following the typical shape of the KE SEE. As expected, the SEEs for the KE linear are U-shaped (i.e., they are relatively large at the ends of the score range and small in the middle), and they are smaller for the linear equating than for the KE equipercentile equating; the SEEs range from 0.50 (at Score 0) to 0.09 (from Score 40 to 42).

**Table 6**

***The PRE Values for the KE Equipercentile (With Optimal Bandwidths) and KE Linear (With Large Bandwidths) for Equating X to Y in the Combined Group (EG Design)***

| $p^{th}$ moment | PRE (KE equip.) | PRE (KE linear) |
|---|---|---|
| 1 | -0.0097 | 0.0002 |
| 2 | –0.0560 | –0.0013 |
| 3 | –0.1566 | –1.0403 |
| 4 | –0.3217 | –3.2003 |
| 5 | –0.5577 | –6.3477 |
| 6 | –0.8695 | –10.2664 |
| 7 | –1.2612 | –14.7432 |
| 8 | –1.7376 | –19.5860 |
| 9 | –2.2998 | –24.6351 |
| 10 | –2.9542 | –29.7610 |

The SEEs available for the classical linear equating method in the equivalent groups design are about the same size over the score range as those for the KE linear (with large bandwidths); the SEEs for the classical equipercentile equating method are not available.

Because we are ignoring the correlation between $X$ and $Y$ by treating the data from the combined group as an EG design because of software limitations, all SEEs are inflated.

## The Anchor Equatings: Description

After investigating the fit of the smoothed distributions based on preserving five moments for each of the marginal distributions and one through four cross-moments of the observed bivariate distribution, we chose the loglinear model that preserved five moments for each of the two marginal distributions, of $X$ and of $A$, and four cross-moments of the bivariate distribution of $(X, A)$. Similarly, we chose the loglinear model that preserved five moments for each of the two marginal distributions, of $Y$ and of $A$, and four cross-moments of the bivariate distribution of $(Y, A)$.

The optimal bandwidths, or continuization values, for the KE poststratification (KE PSE) are $h(X) = 0.5574$ and $h(Y) = 0.6270$. For the KE PSE linear (i.e., the KE poststratification with large bandwidths), the bandwidths were: $h(X) = 56.90$ and $h(Y) = 62.90$.

The KE chained equating was computed using the stand-alone KE software (ETS, 2004b). In the chained equating, the equipercentile equating function from (2) is a mathematical composition of two linking functions, from $X$ to $A$ on $P$ and from $A$ to $Y$ on $Q$. Therefore, four distributions need to be continuized. The optimal continuization values are $h(X) = 0.55971$, $h(A_P) = 0.57845$ and $h(Y) = 0.62010$, $h(A_Q) = 0.55818$. For the KE CE linear (i.e., the KE chained with large bandwidths), the four bandwidths were all set to 120.

### *Diagnosis of the KE Functions*

The PRE values for the KE PSE equipercentile (with optimal bandwidths) and KE PSE linear (with large bandwidths) are given in Table 7. The PRE values indicate a good match between the KE PSE equipercentile equating function computed at the discrete values of $X$ and the targeted distribution of $Y$ via the external anchor, $A$. The PRE values for the KE PSE linear are larger than those for the KE PSE equiperecentile, indicating a less optimal match between the two discrete distributions.

The PRE values are not available for the KE chained equating. See von Davier et al. (2004) for a discussion about the chained equating.

### *Standard Error of Equating*

The SEEs for the KE PSE with optimal bandwidths range from 1.82 (at Score 0) to 0.10 (at Score 39) following a typical shape of the KE SEE. The SEEs for the KE PSE linear (i.e., KE PSE with large bandwidths) are again U-shaped and smaller than those for the nonlinear equating; the SEEs range from 0.53 (at Score 0) to 0.08 (from Score 35 to 39).

The SEEs available for the other linear equating methods (the Tucker and Levine observed-scores methods) are of similar size over the score range as those for the KE linear; the SEEs for frequency estimation are not available. The reported SEEs for the chain equipercentile have large values at the lower score range (5.88 at Score 4, 5.64 at Score 0) and 0 value at highest scores (Score 44).

Given software limitations, the SEEs are presently not available for both chained methods.

**Table 7**

***The PRE Values for the KE PSE Equipercentile (With Optimal Bandwidths) and KE PSE Linear (With Large Bandwidths) for Equating X to Y in the NEAT Design***

| $p^{th}$ moment | PRE (KE equip.) | PRE (KE linear) |
|---|---|---|
| 1 | –0.0131 | 0.0001 |
| 2 | –0.0619 | –0.0009 |
| 3 | –0.1632 | –0.9146 |
| 4 | –0.3277 | –2.8481 |
| 5 | –0.5629 | –5.7027 |
| 6 | –0.8748 | –9.3034 |
| 7 | –1.2689 | –13.4687 |
| 8 | –1.7503 | –18.0300 |
| 9 | –2.3236 | –22.8419 |
| 10 | –2.9932 | –27.7822 |

## Equating Results

As mentioned before, this research focuses on two types of comparisons for the KE: the comparison of the results of all of the anchor equatings, both traditional and KE versions, with the results of the equating criterion (i.e., the equipercentile equating in the EG design) and the comparison of the results of the traditional equating methods with the KE versions in both the EG and the NEAT designs.

Tables 8 and 9 and Figures 1 and 2 address the first type of comparisons mentioned above: the comparison of the anchor equatings with the EG equipercentile criterion. Tables 10 and 11 address the second type of comparison: how well KE approximates the methods it is supposed to approximate (with both sets of methods in the EG and NEAT designs).

**Table 8**

***Equated Scores Corresponding to Selected Raw Scores by Each Equating Method***

| Raw score on Form *X* | 25 | 30 | 35 | 40 |
|---|---|---|---|---|
| Percentile rank (examinees in *P*) | 6th | 16th | 38th | 76th |
| Corresponding score on Form *Y*, as determined by | | | | |
| Chained linear | 14.21 | 20.23 | 26.24 | 32.25 |
| Kernel, chained, large bandwidth | 14.52 | 20.42 | 26.31 | 32.22 |
| Tucker | 15.01 | 20.82 | 26.63 | 32.44 |
| Kernel, poststratfication, large bandwidth | 14.78 | 20.70 | 26.62 | 32.54 |
| Levine observed-score | 14.00 | 20.07 | 26.14 | 32.22 |
| Chained equipercentile | 15.92 | 19.78 | 24.92 | 32.60 |
| Kernel, chained, optimal bandwidth | 15.90 | 19.78 | 24.92 | 32.61 |
| Frequency estimation equipercentile | 16.47 | 20.24 | 25.34 | 32.86 |
| Kernel, poststratfication, optimal bandwidth | 16.44 | 20.25 | 25.35 | 32.86 |
| Criterion equating | 16.09 | 19.81 | 24.98 | 32.93 |

Table 8 presents a comparison of the equated scores produced by each of the anchor equating methods investigated and by the equating criterion for each of four selected raw scores on Form *X*. The rows of the table have been ordered so that each version of KE appears immediately below the (nonkernel) equating method that its results were expected to approximate closely. As expected from the theory and from the previous discussion, the kernel equating results were very close to those of the corresponding nonkernel equating methods—close enough that the difference would not be perceptible on a graph.

Table 9 compares the accuracy of the anchor equating methods investigated—the difference between the equated score as determined by each anchor equating and by the criterion equating—at the four selected raw scores on Form *X*. The curvilinearity of the criterion equating limits the accuracy of the linear equating methods. Even with this limitation, the Levine method performed well. This is not surprising given that the tests and the anchor were carefully constructed and the tests had the same length/reliability (see Petersen et al., 1982). The same information given in Table 9 is plotted in Figure 1.

**Table 9**

***Difference Between Each Anchor Equating and Criterion Equating for Selected Raw Scores***

| Raw score on Form *X* | 25 | 30 | 35 | 40 |
|---|---|---|---|---|
| Difference from criterion equating | | | | |
| Chained linear | –1.88 | 0.42 | 1.26 | –0.68 |
| Kernel, chained, large bandwidth | –1.57 | 0.61 | 1.33 | –0.71 |
| Tucker | –1.08 | 1.01 | 1.65 | –0.49 |
| Kernel, poststratfication, large bandwidth | –1.31 | 0.89 | 1.64 | –0.39 |
| Levine observed-score | –2.09 | 0.26 | 1.16 | –0.71 |
| Chained equipercentile | –0.17 | –0.03 | –0.06 | –0.33 |
| Kernel, chained, optimal bandwidth | –0.19 | –0.03 | –0.06 | –0.32 |
| Frequency estimation equipercentile | 0.38 | 0.43 | 0.36 | –0.07 |
| Kernel, poststratfication, optimal bandwidth | 0.35 | 0.44 | 0.37 | –0.07 |

Figure 2 reflects a more detailed comparison of the nonlinear anchor equatings with the equating criterion. Figure 2 plots the equating differences between the anchor equatings available and the equating criterion; Figure 2 also gives information about the differences among the equating results at the extreme score ranges, while Table 9 and Figure 1 focus on the score ranges where the data are. Again, one can see that, both in the PSE and in the CE case, KE is doing better than or similar to the methods it is supposed to approximate. In all instances, the differences between the equating in question and the criterion equating are less than a half point, which means the differences are smaller than a difference that matters (DTM; Dorans & Feigenbaum, 1994); the DTM depends on the reporting scale and style, and in most cases it is considered to be a half point on the raw scale.

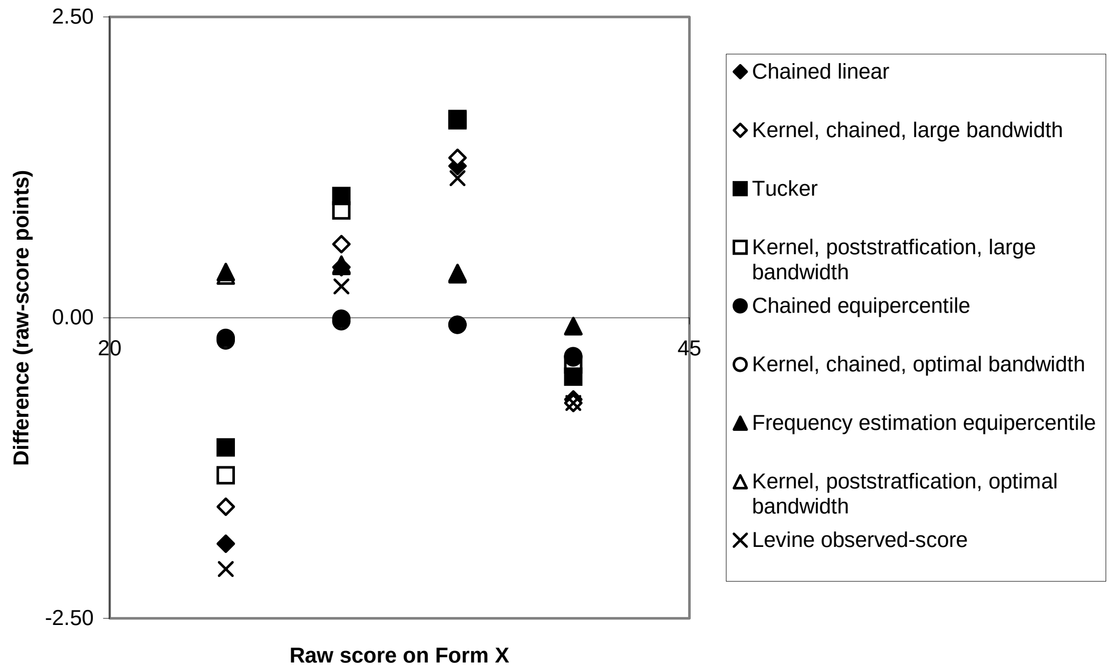


*Figure 1.* **Differences from criterion equating.**

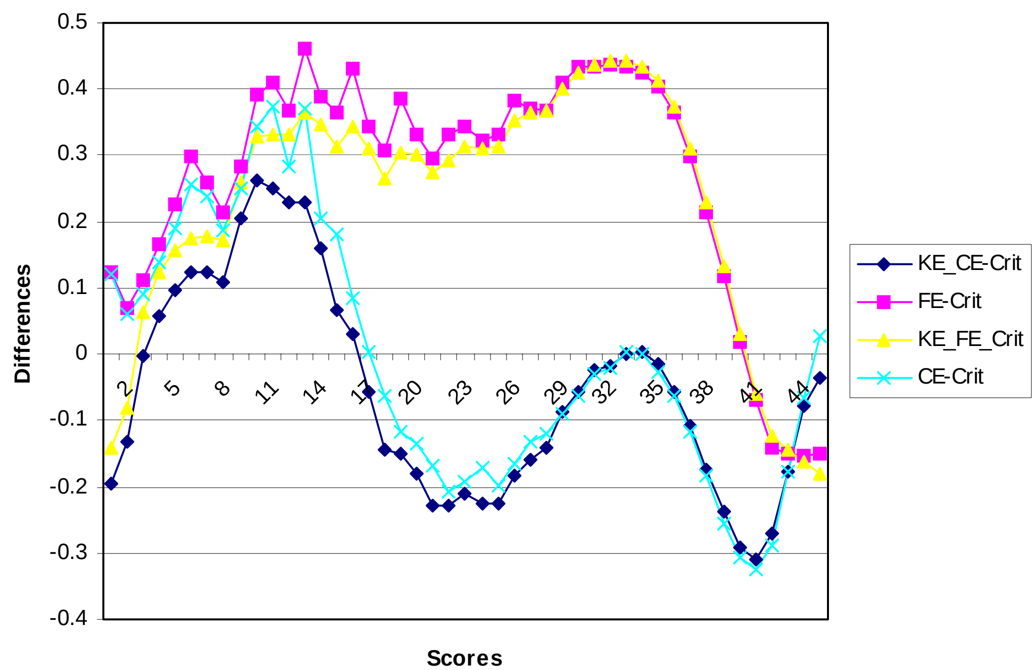


*Figure 2.* **Equating comparisons with the criterion.**

In the rest of this section, we discuss comparisons between the KE methods and those traditional methods that the KE supposedly approximates. In order to compare the results across methods in the NEAT design and with the equating criterion from the EG design, we also computed (a) differences between the equating functions at each corresponding score, (b) the maximum, minimum, and averages of these differences, and (c) the root mean expected error of these differences (reported in Tables 10 and 11). The root mean square difference (RMSE), or error, is

$$\mathrm{RMSE} = \sqrt{\bar{d}^{\,2} + sd_d^{\,2}}\,, \tag{4}$$

where $\bar{d}$ is the mean of the differences of the equated scores ($d = a_i - b_i$, where $a_i$ and $b_i$ denote the equated scores of the score $x_i$ by two different methods, respectively) and $sd$ is the standard deviations of these differences.

Table 10 shows the comparisons of the KE results with the criterion equating in the combined group, as well with the two other nonlinear traditional equatings in the NEAT design.

Table 11 shows the comparisons of the KE linear results with the linear equating method in the combined group and with the Tucker and chained linear methods in the NEAT design.

Tables 8, 9, and 10, which show the equating results, indicate that the KE with optimal bandwidths gives indeed almost identical results to the classical equipercentile equating in the EG design. There are tiny differences at the lowest scores. Also, in the score range where most of the examinees scored, the linear and nonlinear methods are close. As expected, given the large differences in the difficulty of the two test forms, the linear methods provide equated raw scores outside the score range (below zero, on a test where zero is the lowest possible score and a person can expect to get a score of 11 without even reading the questions—that is, by random guessing when the items are all multiple-choice with five options). The KE linear is very close to the classical linear equating in the EG (see the first column in Table 11).

Tables 8, 9, and 10 and Figures 1 and 2 show that the KE PSE with optimal bandwidths gives very similar results to frequency estimation. These tables and figures also show that in the score range where most of the examinees scored, all the methods seem to agree; the nonlinear methods show the same trend at the lower and higher score range. KE PSE linear is, as expected, the closest to the Tucker equating; however, it is not identical with it (see the second column in Table 11). The Levine and Tucker functions seem to differ the most; chained linear is somewhere between the Tucker and Levine functions. Note that the Braun and Holland linear method is not available outside the KE framework (see also Table 2).

**Table 10**

***Summary Measures of Differences Between Nonlinear KE and Its Target Approximations***

| Summary | (EG)<br>Crit—KE | (NEAT)<br>FE—KE PSE | (NEAT)<br>CE-KE CE |
|---|---|---|---|
| Mean diff. | 0.035 | 0.034 | 0.049 |
| SD diff. | 0.053 | 0.060 | 0.065 |
| Max diff. | 0.264 | 0.333 | 0.318 |
| Min diff. | –0.016 | –0.022 | 0.016 |
| RMSE | 0.062 | 0.069 | 0.106 |

Tables 8, 9, and 10 and Figures 2 and 3 show that the KE CE with optimal bandwidths gives similar results to chained equipercentile equating. Also, in the score range where most of the examinees scored, all the methods seem to agree; the nonlinear methods show the same trend at the lower and higher score range. KE CE linear is close to the classical chained linear equating (see the third column in Table 11).

**Table 11**

***Summary Measures of Differences Between Linear KE and Its Target Approximations***

| Summary | (EG)<br>Lin–KE lin. | (NEAT)<br>Tuck–KE PSE | (NEAT)<br>CE lin.–KE CE |
|---|---|---|---|
| Mean diff. | –0.013 | 0.284 | –0.35649 |
| SD diff. | 0.016 | 0.277 | 0.284116 |
| Max diff. | 0.001 | 0.738 | 0.11921 |
| Min diff. | –0.052 | –0.189 | –0.8214 |
| RMSE | 0.020 | 0.397 | 0.45586 |

The differences reported in Tables 10 and 11 indicate a very good match among the methods. Moreover, given the information about the SEES described above (which, for the KE, optimally range between 2.0 and 0.8 for most of the cases), the observed differences are all at (about) the noise level (i.e., the level of uncertainty reflected by the SEEs).

## Conclusions

This study evaluates the kernel method of test equating (Holland & Thayer, 1989; von Davier et al., 2004a) in special settings where an equating criterion is available. The evaluation consists of two aspects: the KE results are compared with those equating results obtained using classical equating methods that the KE functions claim to approximate and the comparisons of the equating results with the equating criterion (which was defined to be the equipercentile equating function in the combined group).

This analysis takes place in a NEAT design with external and internal anchors (the results for the internal anchor are provided in Appendix C) and in an EG design (the combined group in this case). To obtain the data for the NEAT design and for the equating criterion, we constructed pseudotests with real data.

The KE functions agree well with those anchor equating functions that it is supposed to approximate (and that were available for comparisons). The results for the internal anchor are similar to those for the external anchor. Similarly, the results indicate that the KE (both linear and nonlinear) and the equating functions (linear and nonlinear) from the EG design (combined group) agree very well over the whole score range.

Moreover, the comparisons among equating functions versus the equating criterion indicate that the KE results are in most cases closer to the criterion than the other equating functions.

In addition, the KE method provides accuracy measures that are not available for other equating methods. For instance, the percent relative error has already been implemented in the software available for the KE PSE. The other important accuracy measures, such as the standard error of equating difference, will be available in the newly developed KE software.

Based on these results, we recommend that KE be used operationally together with the other equating methods that are usually computed. In addition, careful comparisons of the KE method with other methods and various testing steps of the software in various operational settings should be continued.

The research on the KE method should extend to the investigation of equating trends. This can be done by applying the KE method on data sets that span several years and come from assessment programs that have done equating on a regular basis using classical observed-score methods and then comparing the KE results with the results from the other methods.

# References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.

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., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus post-stratification 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*(1).

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright, *Technical issues related to the introduction of the new SAT® and PSAT/NMSQT®* (RM-94-10). Princeton, NJ: ETS.

ETS. (2004a). GENASYS [Computer software]. Princeton, NJ: Author.

ETS. (2004b). KE software, version 0.1 [Computer software]. Princeton, NJ: Author.

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

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS RR-89-07). Princeton NJ: ETS.

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.

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

Livingston, S. A. (1993). *An empirical tryout of kernel equating* (ETS RR-93-33). Princeton, NJ: ETS.

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

Moses, T., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (ETS RR-04-27). Princeton, NJ: ETS.

Petersen, N. S., Marco G. L, & Stewart, E. E. (1982). A test of adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Macmillan.

## Appendix A

## An Outline of Kernel Equating

This appendix gives a brief outline of the kernel method of observed score test equating or KE (see also von Davier, Holland, & Thayer, 2004a, Appendix B). The KE method is discussed in detail in von Davier, Holland, and Thayer (2004a) for all of the standard equating designs. As mentioned before, the KE method has five basic steps.

### *Step I: Presmoothing*

In this step, the data that are collected in an equating design are presmoothed using standard statistical procedures designed to estimate the actual score distributions that arise in the equating design. Presmoothing, using various techniques, has become a standard tool in various approaches to equipercentile equating.

We advocate using loglinear models for univariate and bivariate score distributions, as discussed in Holland and Thayer (2000), because of their extreme flexibility and ability to accommodate the many unusual features of score distributions that arise in practice. The results of this presmoothing process are twofold. First, the smoothed score distributions that are needed for the rest of the equating process are obtained, and second, a matrix that can be used to calculate the standard error of equating later on in the process is computed. Every presmoothing method has such a matrix, but the loglinear methods have a standard way of finding it in an efficient manner. This is discussed in detail in Holland and Thayer (2000).

### *Step 2: Estimating Score Distributions for the Target Population*

Once the presmoothing has been done and depending on the equating design, formulas are employed that use the smoothed score distribution estimates to produce estimates of the score probability distributions on $T$, which we call $\boldsymbol{r}$ and $\boldsymbol{s}$, where

$$r_j = P\{X = x_j|T\},\ s_k = P\{Y = y_k|T\} \tag{A1}$$

and the vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ are given by

$$\boldsymbol{r} = (r_1, \ldots, r_J), \text{ and } \boldsymbol{s} = (s_1, \ldots, s_K). \tag{A2}$$

The score probabilities for $X$ are associated with the $X$-raw scores, $\{x_j\}$, and those for $Y$ are associated with the $Y$-raw scores, $\{y_k\}$. Depending on the equating design the score probabilities,

***r*** and ***s***, are computed through the design function, which ranges from the simple identity function to the complexities implicit in anchor test methods.

***Step 3:Continuizing the Discrete Score Distributions***

This step is often overlooked in discussions of equipercentile equating methods, but it occurs in all of them. We start with *discrete* score distributions for $X$ and $Y$ on $T$ and turn these into *continuous* score distributions over the whole real line. It is similar to approximating the probabilities from the *discrete* binomial distribution by probabilities from the *continuous* normal distribution. Thus, it is a step that looks like an everyday statistical method, but it is actually unusual because the entire discrete distribution is changed into a continuous one that is close to the original in a sense that is often left vague. Our approach is to make this step explicit and to make the sense of the approximation clear. Other equipercentile equating methods replace the discrete score distributions by piecewise linear cdfs based on percentile ranks. The (Gaussian) kernel method of continuizing ***r*** uses the formula

$$F_T(x;\, h_X) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_{XT}}{h_X a_X}\right), \tag{A3}$$

where, $\mu_{XT} = \sum_j x_j r_j$, $\sigma^2_{XT} = \sum_j (x_j - \mu_{XT})^2 r_j$, and $a_X = \sqrt{\sigma^2_{XT} / (\sigma^2_{XT} + h_X^2)}$.

$\Phi(z)$ denotes the standard $N(0, 1)$ cdf, $x$ ranges over $(-\infty, +\infty)$, and $h_X > 0$. $F_T(x;\, h_X)$ is the continuized cdf based on the discrete score distribution determined by ***r*** and $\{x_j\}$. $\mu_{XT}$ and $\sigma^2_{XT}$, given above are the moments of $X$ on $T$.

The continuized $G_T(y,\, h_Y)$ is computed in a similar way using the score probabilities from ***s***, and the $Y$-scores, $\{y_k\}$.

An essential feature of Gaussian kernel continuization is the choice of the bandwidths, $h_X$ and $h_Y$. We recommend using a penalty function to select the bandwidths automatically to make the density functions, $f_T(x;\, h_X)$ and $g_T(y,\, h_Y)$, derived from $F_T(x;\, h_X)$ and $G_T(y,\, h_Y)$ both smooth and able to track the essential features of the smoothed discrete score probabilities. We have found the following penalty functions to give good results:

$$\text{PENALTY}_1(h) = \sum_j \left[(r_j/d_j) - f_T(x_j;\, h)\right]^2, \tag{A4}$$

where $d_j$ is the width of the interval associated with the score $x_j$ (often these widths are all set equal to 1).

$$\text{PENALTY}_2(h) = \sum_j A_j(1 - B_j)\,, \tag{A5}$$

where $A_j = 1$ if the derivative of $f_T(x; h)$ with respect to $x$, $u(x; h)$, is less than 0 a little to the left of $x_j$, and $B_j = 0$ if $u(x; h) > 0$ a little to the right of $x_j$. Thus, we get a penalty of 1 for every score point where the density $f_T(x; h)$ is U-shape around it. What near means is a parameter of $\text{PENALTY}_2(h)$, and we can combine the two penalties with a weight, that is,

$$\text{PENALTY}_1(h) + K{*}\text{PENALTY}_2(h). \tag{A6}$$

We have found $K = 1$ to be useful in several applications where there are teeth or gaps in the distribution that need to be smoothed out. Standard derivative-free methods can be used to minimize these penalty functions in order to choose $h$. Separate continuizations of the two discrete score distributions are carried out, resulting in $F_T(x; h_X)$ and $G_T(y; h_Y)$.

***Step 4: Computing and Diagnosing the Equating Function***

Once all the above work is done, the KE equipercentile equating function can be computed directly as the function composition:

$$e_{XY}(x) = G_T^{-1}(F_T(x; h_X); h_Y) \tag{A7}$$

where $G_T^{-1}(p; h_Y)$ denotes the inverse of $p = G_T(y; h_Y)$. $e_{XY}(x)$ is designed to exactly match the two continuized distributions, but we really want to know how well it does for the discrete distributions. What is important about an equating function is how well $e_{XY}$(X)—as the function of the discrete $X$—matches the discrete target-distribution, $Y$. In order to assess this match, we compare up to the 10th moment of the two distributions, $e_{XY}(X)$ and $Y$, via the percent relative error in the pth moment (PRE) formula: Let

$$\mu_p(Y) = \Sigma_k\,(y_k)^p s_k, \text{ and } \mu_p(\mathrm{e_Y}(X)) = \Sigma_j\,(\mathrm{e_Y}(x_j))^p r_j\,, \text{ then}$$

$$\text{PRE(p)} = 100\text{x}[\mu_p(\mathrm{e_{XY}}(X)) - \mu_p(Y)]/\,\mu_p(Y). \tag{A8}$$

### *Step 5: Computing the Standard Error of Equating and Related Accuracy Measures*

The standard error of equating (SEE) for $e_{XY}(x)$ depends on three factors that correspond to the above four steps—presmoothing, computing $\boldsymbol{r}$ and $\boldsymbol{s}$ from the smoothed data, and the combination of continuization and the mathematical form of the equating function from Step 4. Being based on analytical formulas, KE allows us to use the Taylor expansion or delta method to compute the SEE for a variety of equating designs. The main difference between the various equating designs, as far as computing the SEE for KE is concerned, is Step 2. Each design requires a different formula (a design function) for mapping the presmoothed data to the score probabilities, $\boldsymbol{r}$ and $\boldsymbol{s}$, but the contributions of the other steps to the SEE are the same for *all* designs. This observation allows a general computing formula for the SEE to be devised for KE that reflects presmoothing, the equating design, and the use of Gaussian kernel smoothing for continuizing the discrete cdfs. The standard error of equating difference (SEED) discussed in von Davier et al. (2004) is a new tool, unique to KE, for evaluating the degree to which KE and linear equating agree. Moreover, the SEED can be used to assess if the difference between two equating functions that share the same parameters are statistically significant.

## Appendix B

### Equated Scores Corresponding to Selected Raw Scores

| Raw score on Form *X* | 25 | 30 | 35 | 40 |
|---|---|---|---|---|
| Percentile rank (examinees in *P*) | $6^{th}$ | $16^{th}$ | $38^{th}$ | $76^{th}$ |
| Corresponding score on Form *Y*, as determined by | | | | |
| Criterion equating (equating *X* to *Y* in the combined group) | 16.09 | 19.81 | 24.98 | 32.93 |
| Equating *X* to *Y* in *P* | 16.33 | 19.98 | 24.96 | 32.92 |
| Equating *X* to *Y* in *Q* | 15.54 | 19.46 | 25.00 | 33.05 |

*Note.* The traditional equipercentile equating method was used for equating *X* to *Y* in the combined group (the criterion equating) and in each of the separate groups, *P* and *Q*.

The differences reported here are small for those score points where we have data and seem to be at the noise level (given the SEE). The difference between the equating results at the lower score points seems to be significant.

As mentioned before, this check of the invariance of the equating criterion is not for the purpose of validating the choice of the criterion.

# Appendix C

## Brief Description of the Results for the NEAT Design With an Internal Anchor

As in the case of the external anchor, the research on the NEAT design with an internal anchor focuses on two types of comparisons for the KE: the comparison of the results of all of the anchor equatings, both traditional and KE versions, with the results of the equating criterion (i.e., the equipercentile equating in the EG design) and the comparison of the results of the traditional equating methods with the KE versions in both the EG and the NEAT designs.

This appendix presents three tables that contain the comparison results among the anchor equatings and kernel equating for the internal anchor case. Tables C1 and C2 address the first type of comparisons mentioned above: Compare the anchor equatings with the EG equipercentile criterion. Table C3 addresses the second type of comparison, that is, how well KE approximates the methods it is supposed to approximate (both sets of methods in the EG and NEAT designs). Table C1 shows the comparisons of the KE results with the criterion equating in the combined group, as well comparisons of the criterion equating with the two other nonlinear traditional equatings in the NEAT design. Table C2 shows the comparisons of the traditional nonlinear equating methods with the KE versions in both the EG and the NEAT designs. Table C3 shows the comparisons of the KE linear results with the linear equating method in the combined group and with the Tucker and chained linear methods in the NEAT design.

## Test Construction and Procedure

The anchor has 24 items and is the same as described in the body of the report. The tests to be equated are $X + A$ and $Y + A$, both in the NEAT design with the internal anchor $A$ and in the combined group. This decision was made to allow the scale comparability across the two designs. The unique tests, $X$ and $Y$, have 44 items each and are the same as in the body of the report. The summary statistics for the internal case, both in each of the populations and in the combined group, are given in Appendix D.

The criterion equating was chosen to be the equipercentile equating of smoothed distributions of each of the 68-item tests (Form $X + A$ and Form $Y + A$) as in (2) in the combined group (EG design), that is, on a target population that is a weighted average of the two groups from $P$ and $Q$.

### *The KE Results in the Combined Group for the Internal Anchor Case*

The continuization values for the KE are $h(X + A) = 0.6040$ and $h(Y + A) = 0.6484$. For the KE linear, the large bandwidths were: $h(X + A) = 100.70$ and $h(Y + A) = 89.00$.

### *The KE Results in the NEAT Design for the Internal Anchor Case*

The continuization values for the KE PSE are $h(X + A)= 0.6007$ and $h(Y + A) = 0.6477$. For the KE PSE linear, the large bandwidths were $h(X + A) = 93.40$ and $h(Y + A) = 95.50$.

The KE chained equating was computing using the stand-alone KE software (ETS, 2004b). The optimal continuization values are $h(X) = 0.611$, $h(A_P) = 0.58$ and $h(Y) = 0.64$, $h(A_Q) = 0.56$. For the KE chained linear, the large bandwidths were all set to 120.

The SEEs for the KE PSE with optimal bandwidths range from 2.52 (at Score 1) to 0.11 (from Score 54 to 61) following a typical shape of the KE SEE.

**Table C1**

***Summary Measures of Differences Between Nonlinear KE, Frequency Estimation, Chained Equipercentile, and Criterion Equating, Internal Anchor Case***

| Summary | (NEAT) KE PSE—Crit. | (NEAT) FE—Crit. | (NEAT) CE—Crit. | (NEAT) KE CE—Crit. |
|---|---|---|---|---|
| Mean diff. | 0.053 | 0.083 | –0.053 | –0.059 |
| SD diff. | 0.835 | 0.805 | 0.742 | 0.705 |
| Max diff. | 0.838 | 0.869 | 0.836 | 0.747 |
| Min diff. | –2.647 | –2.477 | –2.447 | –2.422 |
| RMSD diff. | 0.837 | 0.809 | 0.744 | 0.707 |

Tables C1 to C3 indicate that the KE with optimal bandwidths gives indeed almost identical results to equipercentile equating. There are small differences at the lowest scores. Also, in the score range where most of the examinees scored, the linear and nonlinear methods are close. KE linear is, as expected, very close to the linear equating for the whole score range. The diagnostic values indicate a good match between the KE optimal and the target distribution of $Y + A$ and are omitted. As expected, the SEEs for the KE linear are U-shaped and smaller for the linear equating; the SEEs range from 0.64 (at Score 0) to 0.13 (from Score 54 to 60).

**Table C2**

***Summary Measures of Differences Between Nonlinear KE and Its Target Approximations, Internal Anchor Case***

| Summary | (EG)<br>EP—KE | (NEAT)<br>FE—KE PSE | (NEAT)<br>CE-KE CE |
|---|---|---|---|
| Mean diff. | 0.021 | 0.031 | 0.006 |
| SD diff. | 0.038 | 0.048 | 0.229 |
| Max diff. | 0.224 | 0.174 | 0.030 |
| Min diff. | –0.030 | –0.010 | –1.660 |
| RMSD diff. | 0.043 | 0.057 | 0.229 |

As seen in Tables C1 to C3, the KE PSE with optimal bandwidths gives very similar results to frequency estimation. Also, in the score range where most of the examinees scored, all the methods seem to agree; the nonlinear methods show the same trend at the lower and higher score range. KE PSE linear is, as expected, the closest to the Tucker equating; however it is not identical to it. The Levine and Tucker functions seem to differ the most; chained linear is somewhere between the Tucker and Levine functions.

As seen in the tables, the results from the internal anchor are very similar to those for the external anchor presented in the report. The differences reported in Tables C1 to C3 reflect a very good match among the methods. Moreover, given the information about the standard error of equatings (SEEs for the KE optimal range between 2.0 and 0.8 for most of the cases) described above, the observed differences are all at (about) the noise level.

**Table C3**

***Summary Measures of Differences Between Linear KE and Its Target Approximations, Internal Anchor Case***

| Summary | (EG)<br>Lin—KE lin. | (NEAT)<br>*T*—KE PSE | (NEAT)<br>CE—KE CE* |
|---|---|---|---|
| Mean diff. | –0.004 | 0.165 | –0.39 |
| SD diff. | 0.005 | 0.185 | 0.27 |
| Max diff. | 0.000 | 0.471 | 0.28 |
| Min diff. | –0.017 | –0.156 | –0.83 |
| RMSD diff. | 0.006 | 0.248 | 0.471 |

## Appendix D

## Summary Statistics

This appendix contains the summary statistics for the internal anchor case. It also contains the information about the correlation of the test and the anchor in both situations, internal and external cases.

**Table D1**

***Summary Statistics for the Observed Frequencies of X (P), Y (Q), and A (P, Q), Internal Anchor***

| Sample | $P$ ($N$ = 6,168) | | | $Q$ ($N$ = 4,237) | | |
|---|---|---|---|---|---|---|
| Form | $X + A_1$ | $Y + A_1$ | $A_1$ | $A_1$ | $Y + A_1$ | $X + A_1$ |
| Mean | 51.16 | 42.62 | 16.03 | 17.00 | 44.98 | 53.38 |
| SD | 9.34 | 10.31 | 4.19 | 3.85 | 9.55 | 8.04 |
| Skewness | –0.71 | –0.19 | –0.37 | –0.53 | –0.37 | –0.86 |
| Kurtosis | 3.16 | 2.44 | 2.59 | 2.82 | 2.64 | 3.66 |
| Obs. min | 12 | 12 | 2 | 2 | 14 | 20 |
| Obs. max | 68 | 67 | 24 | 24 | 67 | 68 |
| Alpha reliability | 0.8799 | 0.8775 | 0.7510 | 0.7254 | 0.8651 | 0.8543 |

*Note*. Total test length = 68. Anchor test length = 24.

**Table D2**

***Summary Statistics for the Observed Frequencies of X (P + Q), Y (P + Q), and A (P + Q), Internal Anchor***

| Sample | $P + Q$ ($N$ = 10,405) | | |
|---|---|---|---|
| Form | $X + A_1$ | $A_1$ | $Y + A_1$ |
| Mean | 52.06 | 16.43 | 43.58 |
| SD | 8.90 | 4.09 | 10.07 |
| Skewness | –0.80 | –0.45 | –0.27 |
| Kurtosis | 3.39 | 2.68 | 2.50 |
| Obs. min | 12 | 2 | 12 |
| Obs. max | 68 | 24 | 67 |
| Alpha reliability | 0.8278 | 0.7446 | 0.8743 |

*Note*. Total test length = 68. Anchor test length = 24.

**Table D3**

***Correlations Between the Anchor and the Tests***

| | $P + Q$ | $P$ | $Q$ |
|---|---|---|---|
| $(X, A_1)$ | 0.768 | 0.782 | 0.735 |
| $(X+A_1, A_1)$ | 0.922 | 0.925 | 0.916 |
| $(Y, A_1)$ | 0.779 | 0.788 | 0.759 |
| $(Y+A_1, A_1)$ | 0.913 | 0.917 | 0.903 |