Research Report

Comparisons of the Kernel Equating Method With the Traditional Equating Methods on Praxis™ Data

Xia Mao

Alina A. von Davier

Stacie Rupp

Research & Development

October 2006
RR-06-30

Comparisons of the Kernel Equating Method With the Traditional Equating Methods on Praxis™ Data

Xia Mao
University of Iowa, Iowa City

Alina A. von Davier and Stacie Rupp
ETS, Princeton, NJ

October 2006

Abstract

Kernel equating (KE) is a new approach to observed-score equating and is described in detail in von Davier, Holland, and Thayer (2004b). Over the past months, several evaluation studies of KE have been designed and carried out. In this part of the overall evaluation study, we compared the KE method with other equating methods using real data from the program of Praxis Series: Professional Assessments for Beginning Teachers®. The goal of this study was to check how closely the KE results agree with the results from other observed-score equating methods in different operational settings. The equating designs involved include an equivalent-groups (EG) design and two nonequivalent groups with anchor test (NEAT) designs with different sample sizes and different characteristics. We compared KE with the equating methods that were used operationally. The results showed that the differences between KE and the traditional equating methods are very small in the EG design, especially in the linear case. In the two NEAT designs, the KE version of poststratification equating with optimal bandwidths produced close results to its analogue, frequency estimation equipercentile equating, except at the lower score range. The KE linear method yielded very similar results to the Tucker method.

Key words: Kernel equating, observed-score equating, NEAT design

Table of Contents

Page

Appendixes

List of Tables

Page

List of Figures

Page

Introduction

Test equating methods are used widely to produce scores that are comparable across different test forms both within a year and across years. The kernel method of test equating or kernel equating (KE) was introduced by Holland and Thayer (1989) and extensively developed by von Davier, Holland, and Thayer (2004b).

The KE method is a rigorously developed statistical procedure that uses a Gaussian kernel to continuize the discrete test score distributions, whereas the classical equipercentile equating method uses a linear interpolation approach to achieve continuous distributions. As a consequence, the KE function, which is a differentiable function, is not expected to completely agree with the classical equipercentile equating function, which is not differentiable. However, the two methods typically produce results that are very close and, after rounding, they might result in the same equated values. Before implementing a new equating method in an operational setting, it is desirable to first demonstrate that the new equating function closely approximates the results from the traditional procedure or procedures being used operationally.

Fortunately, the KE method holds the promise of both approximating the results of several commonly used equating methods and of introducing the routine calculation of new and previously unavailable measures of statistical accuracy. Moreover, it unifies several classical methods of test equating into a single framework, while providing new statistical information that can be used in the practice of test equating. 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. The mathematical and statistical details of KE are reported by von Davier et al. (2004b) and are briefly described in Appendix A.

This study is the second part of a two-part evaluation of the degree to which KE can approximate the results of other equating methods. In this part of the overall evaluation study, we compared the KE with other methods in several equating designs with different sample sizes and different sample characteristics. The goal of this study was to check how closely the KE results agreed with the results from other observed-score equating methods in different operational settings. We investigated three old data sets from the *Praxis*™ program and report the comparisons of KE with the actual equatings that were used operationally.

Livingston (1993) conducted a previous study to evaluate KE. That resampling study used an equivalent-groups design with an internal anchor test. The anchor-equating design was replicated 50 times for four different, small sample sizes. For each replication of the equating design, three separate KEs differing in the bandwidth parameter *h* were computed and compared with chained equipercentile equating of the smoothed and the unsmoothed distributions (the smoothing was achieved by using log-linear models) in terms of accuracy. The results of Livingston's study showed that the equated scores produced by KE were slightly more accurate than those by the chained equipercentile equating of the smoothed discrete distributions but much more accurate than those produced by the chained equipercentile equating of the observed or unsmoothed distributions. The KE 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 study also showed that KE estimates of the standard errors of equating (SEE) were fairly close to the empirical estimates from the 50 replications.

The rest of this report is structured as follows. First, we briefly introduce the equating methods and the equating designs used in this study. We then describe the specific tests and the equating designs and methods that have been used operationally. After that, we present the results of KE and the comparisons with the operational equating methods. The conclusions are drawn in the last section.

Theoretical Background

In this section, we introduce the observed-score equating methods and the equating designs used in the study. We also describe the KE approximations to the operational equating methods.

Observed-Score Equating Methods

There are two test forms to be equated, *X* and *Y*, and a target population, *T*, on which this is to be done. Many observed-score equating methods are based on the equipercentile equating function. It is defined on the target population, *T*, as shown in Equation 1:

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

where $F_T(x)$ and $G_T(y)$ are the cumulative distribution functions (cdfs), of *X* and *Y*, respectively, on *T*. In order for this definition to make sense, we also assume that $F_T(x)$ and $G_T(y)$ have been made continuous or continuized, so that the inverse functions exist for $F_T(x)$ and $G_T(y)$.

Several important observed-score equating methods may be viewed as only differing in the way that the continuization is achieved. The traditional equipercentile equating method (percentile rank method) uses linear interpolation of the discrete distribution to make it piecewise linear and therefore continuous. KE uses Gaussian kernel smoothing to approximate the discrete histogram by a continuous density function. Details are given in von Davier et al. (2004b) and briefly mentioned in Appendix A.

Linear equating assumes that $F_T(x)$ and $G_T(y)$ are continuous and have the same shape while differing in mean and variance. The linear equating function, $\mathrm{Lin}_{XY;T}(x)$, is defined by Equation 2:

$$\mathrm{Lin}_{XY;T}(x) = \mu_{YT} + \sigma_{YT}((x - \mu_{XT})/\sigma_{XT}). \tag{2}$$

Theorem 1 of von Davier, Holland, and Thayer (2003; 2004a) showed that any equipercentile equating function can be decomposed into the corresponding linear equating function and a nonlinear part. In the KE framework, linear equating is a special case of equipercentile equating.

Equating Designs

This study involved two major data collection designs, the equivalent-groups (EG) design and the nonequivalent groups with anchor test (NEAT) design.

EG design. The EG design involves two samples of test takers, randomly and independently drawn from a population *P*. In this study, the target population in the EG design is considered to be *P*.

The two operational tests to be equated, *X* and *Y*, are given to two samples of examinees from one population, resulting in the data structure shown in Table 1, where ✓ denotes the presence of data.

Table 1

EG Design

Population	Sample	*X*	*Y*
P	G_1	✓	
P	G_2		✓

The non-equivalent groups with anchor test design. The NEAT design involves two populations of test takers, *P* and *Q* (usually different test administrations), and makes use of an anchor test to link the two tests. The two operational tests to be equated, *X* and *Y*, are given to two samples of examinees from the two populations or administrations. In addition, an anchor test, *A*, is given to both samples from *P* and *Q*, as shown in Table 2. The anchor test, *A*, can be either a part of both *X* and *Y* (the internal anchor case) or a separate test (the external anchor case).

Table 2

NEAT Design

Population	Sample	*X*	A	*Y*
P	G_1	✓	✓	
Q	G_2		✓	✓

The target population, *T*, for the NEAT design is a weighted average of *P* and *Q*. *P* and *Q* are given weights that sum to 1. This is denoted by Equation 3:

$$T = wP + (1 - w)Q \tag{3}$$

The partition of *T* is determined by the weight *w* (see Angoff, 1971/1984; Kolen & Brennan, 2004; or von Davier et al., 2004a, for a discussion of the target population in the NEAT design and of the role of the weights).

In the NEAT design, the two most important test scores, *X* and *Y*, are each only observed on either *P* or *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 in order to overcome this lack of complete information in the NEAT design. The basic task for any equating method in the use of the NEAT design is—under acceptable and sufficiently strong assumptions—to obtain the (moments of the) distributions of *X* and *Y* on *T*.

The two classes of observed-score equating methods used in the NEAT design that concern us here include chained equating (CE), including the equipercentile and linear methods), and poststratification equating (PSE), including the frequency estimation equipercentile (FE) method, the Tucker method, and the Braun and Holland (1982) linear method. The details of these methods have been described by Kolen and Brennan (2004).

KE Methods

KE methods in the EG design. In the EG design, the target population is simply the population from which the two groups are sampled, and the two samples provide data that can be used to estimate score probabilities on the target population. The traditional observed-equating methods for this design include linear equating and equipercentile equating (Kolen & Brennan, 2004). Linear equating uses a linear relationship to describe the form-to-form differences in difficulty. Equipercentile equating uses a curve to describe the form-to-form differences in difficulty.

The KE method is based on the equipercentile definition of equating, and linear equating is treated as a special case within the framework. KE uses Gaussian kernel smoothing to approximate the discrete score distribution by a continuous density function. By varying the bandwidth values (see Step 4 in Appendix A), KE can approximate the equipercentile and linear equating methods. When optimal bandwidths are chosen, KE will approximate the traditional equipercentile equating method. When the bandwidths used are 10 times the standard deviation of the scores or larger (i.e., large bandwidths), the continuized distributions will be nearly normal, in which case the KE functions can be regarded as approximately linear. Thus, linear equating can be regarded as a special case of equipercentile equating in the framework of the KE.

KE methods in the NEAT design. In the NEAT design, a variety of equating methods is available. In this study, we are concerned with three general cases: (a) PSE, (b) CE, and (c) Levine linear methods.

PSE methods include the frequency estimation, the Tucker observed-score linear method, and the Braun and Holland or PSE-linear method (Kolen & Brennan, 2004; von Davier et al., 2004b). KE methods include approximations to the frequency estimation and the PSE-linear methods. When optimal bandwidths are chosen to closely approximate the discrete distribution, then the KE version of PSE (KE PSE) will approximate frequency estimation. When large bandwidths are chosen (KE PSE-linear), the result will approximate the PSE-linear method. The PSE-linear method uses the poststratified score probabilities to compute the mean and variance of X and Y on T. These moments then are used to directly compute $\mathrm{Lin}_{XY;T}(x)$, defined in Equation 2. This approach makes different assumptions than does the Tucker linear method, though they are related. In the Tucker linear equating, the assumptions regarding the population invariance are weaker because they refer only to the first two moments of the conditional distributions of X and Y. However, the additional assumptions, linear regression and constant conditional variance, are stronger in the sense that linear PSE can have nonlinear regressions of X and Y on A as well as nonconstant conditional variances. These conditional moments depend on the data and on the model used to presmooth the bivariate data tables that arise in the NEAT design. Hence, the Tucker linear method and the PSE-linear method will agree only in specific circumstances, when the regressions of *X* and *Y* given *A* are linear enough and the conditional variances are almost constant.

The second case of observed-score equating with the NEAT design involves the CE methods. These include both the chained linear and chained equipercentile methods. These, too, may be approximated by KE. The KE version of CE (KE CE) will approximate the chained equipercentile method when the small bandwidths are used and will approximate the chained linear method (KE CE Lin) when large bandwidths are used.

The third case is the Levine linear observed-score method. This is a linear method with no known curvilinear analogue. It can be justified by the use of classical test theory. However, in the present state of development, KE does not have a version that approximates the Levine method.

Thus, KE should be able to come close to reproducing the usual equipercentile and linear methods for both PSE and CE. In our study, there are certain limitations on all the comparisons that could be made that are implied by the above discussion, due to limitations of the software that we used—GENASYS (ETS, 2004a). For example, GENASYS cannot compute the PSE-linear method. Thus, we only could compare the KE with the FE method and the Tucker method using GENASYS. A newly developed computer software, KE-Software (ETS, 2004b), will be able to perform computation for all the KE methods.

Data Description

This study used real data from the *Praxis Series: Professional Assessments for Beginning Teachers*®, which is a set of tests designed for use by state education agencies in making teacher licensing decisions. The series covers different subject areas such as mathematics, biology, arts, and general science. For each test, different equating designs may be used. Typically, several equating methods are conducted operationally for each test. In this study, we applied KE and compared it with the other methods that were used operationally in different equating designs. The following section gives a description of the data, including the tests and the equating designs used in this study.

EG Design

The data for the EG design came from a recent administration of the *Praxis Series: Pre-Professional Skills Test in Mathematics*. The *Praxis Series: Pre-Professional Skills Test in Mathematics* is designed to measure the mathematical skills and concepts that an educated adult might need. It is a 1-hour test with 40 multiple-choice questions scored as right or wrong. Each item is classified into one of three content areas, and the proportions of the three content areas range from 25% to 45% of the total test. The two samples of examinees from the same population were obtained by the spiraling method. Group 1 took test form *X* and Group 2 took test form *Y* (see Table 1 for details). Since we know that the spiraling was carried out carefully, we conjecture that the difference in the samples might be due to how the spiraling was implemented in various batches or centers or to exclusions of test takers for various reasons (leaving the room in first minutes or empty papers, for example).

Table 3 shows the summary statistics for the two tests being equated. From Table 3, we can see that the new form is slightly more difficult than the old form.

Table 3

Summary Statistics for the Two Test Forms in the EG Design

Statistic	New form (*X*)	Old form (*Y*)
N	2,968	3,017
Mean	24.92	25.23
Median	24.75	25.44
SD	7.02	7.57
Skewness	–0.0102	–0.1636
Kurtosis	2.31	2.22
Min (Possible min)	6.00 (0.00)	5.00 (0.00)
Max (Possible max)	40.00 (40.00)	40.00 (40.00)

In all the operational equatings that have been conducted, the data were smoothed using log-linear models (see Holland & Thayer, 2000), as in an EG design. The log-linear model used to fit a univariate distribution is Equation 4:

$$\log_e(p_j) = \alpha + \sum_{i=1}^{I} \beta_i (x_j)^i , \qquad (4)$$

where the $(x_j)^i$ is a vector of score functions, α is as normalized constant, and β_i is the *I* free parameter to be estimated in the model-fitting process (see Holland & Thayer, 1987, 2000; Moses, von Davier, & Casabianca, 2004, for details on the log-linear models.) The value of *I* determines the number of moments of the actual test score distribution that are preserved in the fitted distribution. If $I = 1$, then the fitted distribution preserves the first moment (the mean) of the observed distribution. If $I = 4$, then the fitted distribution preserves the first, second, third, and fourth moments (mean, variance, skewness, and kurtosis) of the observed distribution (Holland & Thayer, 2000).

The EG design results in two univariate distributions, denoted by *X* and *Y*. In the operational equating for these tests, a decision was made (prior to this study) to preserve the first five moments for each univariate distribution. Table 4 presents the fit statistics. For these univariate distributions, the likelihood ratio statistic and the Freeman-Tukey (FT) chi-square (global) statistic (Holland & Thayer, 2000) were expected to be chi-squared distributed with the nominal degrees of freedom, *df*, if the model fit the data. Table 4 shows that both the likelihood ratio chi-square and the FT statistics were close to the degrees of freedom. This indicated that the model fit the data well.

Table 4

Summary of the Fit Measures for the Fitted Log-Linear Models in the EG Design

Measure	New form (*X*)	Old form (*Y*)
Moments preserved	5	5
Likelihood ratio	35.62	30.29
df	35	35
FT chi-square	33.50	30.04

Note. FT = Freeman-Tukey.

The FT deviates at each score (Holland & Thayer, 2000) also were examined to investigate fit. The FT deviates behave roughly like independent standard normal deviates if the model fits adequately, so a deviate with an absolute value greater than 3 would be considered large. In this model, the FT deviates scattered randomly around zero, and there were no large FT deviates in either distribution. We also checked the fitted versus observed distribution plots and found that the smoothed frequencies were, in general, close to the observed frequencies. All these indicated a good fit between the raw and smoothed data.

The equating methods that have been conducted operationally include linear equating and equipercentile equating. In this study, we conducted KE with optimal bandwidths and KE with large bandwidths using GENASYS (ETS, 2004a). We then compared the KE results with those from the linear and nonlinear methods used operationally.

NEAT Design With Internal Anchor

The data used for this NEAT design came from two different administrations of the Art: Content Knowledge test. The Art: Content Knowledge test is designed for prospective teachers who have completed teacher training programs. It is a 2-hour test with 120 multiple-choice questions scored as right or wrong, of which 30 are common items. Each item is classified into one of three content areas, and the proportions of the three content areas range from 25% to 39% of the total.

Table 5 shows the summary statistics for the two tests and the anchor. The table shows that the sample means for group *P* and group *Q* on the anchor test *A* were very close. The ratio of the standard deviations on the anchor test in the two groups was 1.01, and the standardized difference of means on the anchor test in the two groups was 0.01, where the standardized difference is calculated as $\frac{|\mu_{AP} - \mu_{AQ}|}{SD_A}$ (note that μ_{AP} and μ_{AQ} are the means on the anchor in population *P* and population *Q*, respectively. SD_A is the averaged standard deviation of the anchor in the two populations). These indicated that the two groups were very similar in their abilities.

In all the operational equatings that have been conducted, the two bivariate discrete observed distributions were smoothed using log-linear models (Holland & Thayer, 2000). Log-linear models were fitted to the two bivariate distributions, (*X* + *A*, *A*) in *P* and (*Y* + *A*, *A*) in *Q*. The log-linear model for fitting the bivariate distributions (*X* + *A*, *A*) in *P* is Equation 5:

$$\log_e(p_{jl}) = \alpha + \sum_{i=1}^{T_{XP}} \beta_{XP_i}(x_j)^i + \sum_{i=1}^{T_{AP}} \beta_{AP_i}(a_l)^i + \sum_{i=1}^{I_{XP}} \sum_{i'=1}^{I_{AP}} \beta_{XAP_{ii'}}(x_j)^i (a_l)^{i'} \ , \tag{5}$$

where p_{jl} is the joint score probability of the score (x_j', a_l) (score x_j' on the test $X + A$ and score a_l on the anchor A). The fitting of the model in Equation 5 produces a fitted bivariate distribution that preserves T_{XP} moments in the marginal (univariate) distribution of $X + A$, T_{AP} moments in the marginal (univariate) distribution of the anchor A, and I_{XP}, I_{AP} cross-moments in the bivariate ($X + A$, A) distribution. This model is also appropriate for the smoothing of bivariate distributions with impossible ($X + A$, A) score combinations, or *structural zeros*. Distributions with structural zeros arise when ($X + A$, A) represent the scores of a total test and an internal anchor, so that the total test score can never be less than the score on the internal anchor test (Holland & Thayer, 2000). The same type of model would be fit to the ($Y + A$, A) frequencies in Q.

Table 5

Summary Statistics for the Two Test Forms and the Anchor in the NEAT Design (Internal Anchor)

	Group *P*		Group *Q*	
Statistic	Test ($X + A$)	Anchor *A*	Test ($Y + A$)	Anchor *A*
N	524	524	733	733
Mean	72.76	19.72	81.57	19.78
Median	72.44	19.73	82.19	19.89
SD	13.43	4.26	13.57	4.24
Skewness	–0.03	–0.12	–0.20	–0.26
Kurtosis	2.58	2.61	2.73	2.79
Min (Poss. min)	38.00 (0.00)	8.00 (0.00)	39.00 (0.00)	7.00 (0.00)
Max (Poss. max)	105.00(112.00)	30.00 (30.00)	114.00 (120.00)	30.00 (30.00)
Reliability	0.89	0.70	0.90	0.72
SEM	4.42	2.34	4.27	2.26
Correlation [a]	0.88		0.89	

[a] Refers to the correlation between the test and the anchor.

In the equating process for these tests, a decision was made prior to this study to fit a (5 + 5 + 1) model for the bivariate distribution of ($X + A$, A) on P. That means the first five moments for the marginal distribution of $X + A$, the first five for that of A, and the correlation of X and A were preserved. The same model was fit to the distribution of *(Y + A, A)* on *Q*. The first and third columns of data in Table 6 show the fit statistics for this model. The likelihood ratio chi-square statistics do not follow chi-square distributions with those nominal degrees of freedom in the table due to the sparseness of data at a large number of score points in the bivariate distributions. As a check on the need for the cross-moment included in the model, we also fit a (5 + 5 + 0) model to both bivariate distributions. The change in the likelihood ratio chi-square between these two models was 773.57 for the new form and 1124.21 for the old form on one degree of freedom. The Akaike information criterion (AIC) and consistent Akaike information criterion (CAIC; Akaike, 1974) are used to evaluate the tradeoff between biased estimates that result from smoothing too much and unstable estimates that result from not smoothing enough. The AIC and CAIC both decreased substantially when the cross-moment was added. These results gave evidence for the need to add the cross-moment to improve the fit.

Table 6

Summary of the Fit Measures for the Fitted Log-Linear Models in the NEAT Design (Internal Anchor)

Measure	New form (*X* + *A*)		Old form (*Y* + *A*)	
Moments preserved	5 + 5 + 1	5 + 5 + 0	5 + 5 + 1	5 + 5 + 0
Likelihood ratio	600.97	1374.54	713.45	1837.66
df	2561	2562	2809	2810
FT chi-square	418.80	792.45	489.97	1143.00
AIC	624.97	1396.54	737.45	1859.66
CAIC	688.11	1454.42	804.61	1921.23

Note. AIC = Akaike information criterion; CAIC = consistent Akaike information criterion; FT = Freeman-Tukey deviate.

We examined the FT deviates to investigate the fit of the marginal distributions of *X* + *A* and *A* in *P* and *Y* + *A* and *A* in *Q*. In this data set, FT deviates behaved roughly like independent standard normal deviates, and there was only one large value (|FT| > 3) of -4.45 at the score point of 61.0 in *Y* + *A*. The plots of fitted versus observed distributions did not show large deviates as well. For a more detailed examination of the fit of the bivariate distributions, we also checked the conditional distributions *X* + *A* given *A* and *Y* + *A* given *A*. The conditional means were nearly linear and were well reproduced by the fitted models. There were some discrepancies between the fitted and observed conditional standard deviations, but the trends of the two were similar. Although other possible models may fit the data better, we decided to keep the (5 + 5 + 1) model for KE, because it fit the data reasonably well, and it made more sense to compare KE with the operational equating methods using the same presmoothing model.

The equating methods that have been conducted operationally include the chained linear equating, the chained equipercentile equating, and the Tucker equating methods. In this study, we computed the FE and compared it and the Tucker method with KE PSE and KE PSE-linear, respectively, using GENASYS (ETS, 2004a).

NEAT Design With External Anchor

The test forms for this NEAT design came from two different administrations of Social Studies: Interpretation of Materials. This test is designed to assess the knowledge and skills for prospective beginning teachers of social studies in a secondary school. It is a 1-hour test with five, two-part, short-answer essay questions. The scoring of each part in each question is based on a 4-point scale (0–3), and the score is the sum of the ratings provided by two raters. Five content areas are covered, and each constitutes 20% of the test. The external anchor is a separate test called Social Studies: Content Knowledge. It is a 2-hour test with 130 multiple-choice questions scored as right or wrong. Table 7 shows that the reliabilities of the tests and the anchors and the correlations between them are relatively low, which means that the equating results are weaker than in the usual circumstances. This is an example of linking of tests coming from a weak collection design. We chose this example on purpose to see how the KE performed in such poor conditions, and we used linking instead of equating for this special case. The two external anchor test forms already had been equated and scaled, and scaled scores on the anchors were used to do the linking discussed here.

Table 7 shows the summary statistics for the two test forms and the anchors. The sample size was relatively small compared with the two tests involving the NEAT design with internal anchor. The sample mean of group P on the anchor test was higher than that of group Q. The ratio of the standard deviation of the anchor tests in the two groups was 0.98, and the standardized difference of the means on the anchor test in the two groups was 0.11. These indicated that the two groups were similar in their abilities.

Table 7

Summary Statistics for the Two Test Forms and the Anchors in the NEAT Design (External Anchor)

	Population (*P*)		Population (*Q*)	
Statistic	Test (*X*)	Anchor (A_1)[a]	Test (*Y*)	Anchor (A_2)[a]
N	100	100	274	274
Mean	56.92	163.61	57.01	161.72
Median	55.70	162.50	57.67	161.25
SD	10.12	16.98	12.43	17.40
Skewness	0.29	0.14	-0.27	0.09
Kurtosis	2.23	2.11	2.68	2.41
Min (Poss. min)	39.00 (0.00)	130.00 (100.00)	23.00 (0.00)	119.00 (100.00)
Max (Poss. max)	82.00 (90.00)	200.00 (200.00)	83.00 (90.00)	200.00 (200.00)
Reliability	0.57	0.72	0.68	0.76
SEM	6.62	9.06	6.98	8.49
Correlation	0.64		0.72	

[a] The scores in the external anchors are scaled scores.

In the operational linking process, a decision was made prior to this study to fit a (3 + 3 + 1) model for both the bivariate distribution of (X, A_1) on *P* and that of (Y, A_2) on *Q*. Columns 1 and 3 in Table 8 that contain data show the fit statistics for this model. The likelihood ratio chi-square statistics do not follow chi-square distributions with those nominal degrees of freedom in the table due to the sparseness of data at a large number of score points in the bivariate distributions. As a check on the need for the cross-moment included in the model, we also fit a (3 + 3 + 0) model to both bivariate distributions. The change in the likelihood ratio chi-square between these two models was 52.36 for the new form and 203.84 for the old form on one degree of freedom. Other indices like AIC and CAIC decreased substantially when the cross-moment was added. These gave evidence for the need to add the cross-moment to improve the fit.

We also examined FT deviates and found only one large value (|FT|>3) of -3.11 at the scale score of 176.0 in A_2. The plots of fitted versus observed distributions showed reasonably good fit as well. The conditional means were nearly linear and were well reproduced by the fitted model.

There were some discrepancies between the fitted and observed conditional standard deviations, but the trends across the two tests were similar. All these indicated that the (3 + 3 + 1) model fit the data reasonably well. For the same reasons as in the NEAT design with internal anchor, we decided to use this smoothing model for the KE in this design.

Table 8

Summary of the Fit Measures for the Fitted Log-Linear Models in the NEAT Design (External Anchor)

Measure	New form (*X*)		Old form (*Y*)	
Moments preserved	3+3+1	3+3+0	3+3+1	3+3+0
Likelihood ratio	627.34	679.70	1251.03	1454.87
df	9183	9184	9183	9184
FT chi-square	194.46	196.08	493.82	510.75
AIC	643.34	693.70	1267.03	1468.87
CAIC	672.18	718.94	1303.93	1501.16

Note. AIC = Akaike information criterion; CAIC = consistent Akaike information criterion; FT = Freeman-Tukey.

The equating methods that have been conducted operationally for this linking case include the chained linear equating, the chained equipercentile equating, and the Tucker equating. In this study, we computed the FE and compared it and the Tucker method with KE PSE and KE PSE-linear, respectively, using GENASYS (ETS, 2004a).

Results

The results of applying the KE method to the three equating designs are presented below. In addition, for each design, comparisons are made between the KE results and the results of its analogues based on the traditional equating methods.

EG Design

KE results. Score distributions were continuized based on the presmoothed distributions that preserve the first five moments. The optimal values of h_x and h_y using Penalty 1 only (see

Appendix A) were 0.6039 and 0.5881, respectively. For the KE linear case, we used bandwidths of 10σ for both distributions (h_x = 70.2, h_y = 75.7).

KE provides a tool called percent relative error (PRE) to investigate how well $\hat{e}_y(X)$, in other words, the KE function at the discrete values of *X* approximates the discrete distribution of *Y* (von Davier et al., 2004). The definition of PRE is provided in Appendix A. PRE (*p*) gives the differences between the p^{th} moments expressed as a percent of the size of the moment of the score being equated. Table 9 shows the PRE values for the KE optimal and KE linear for equating *X* to *Y*. The PRE values for the KE optimal indicated a good match between the equated function computed at the discrete values of *X* and the target distribution of *Y*. The PRE values for the KE linear indicated a poorer match between the linear equating function at the discrete values of *X* and *Y*, which resulted from the different shapes of the two distributions.

Table 9

The PRE Values for the KE Optimal and KE Linear for Equating X to Y in the EG Design

Moments	PRE (optimal)	PRE (linear)
1	–0.0022	0.0000
2	–0.0022	0.0001
3	0.0030	0.3270
4	0.0150	1.1186
5	0.0338	2.3925
6	0.0587	4.1130
7	0.0890	6.2290
8	0.1240	8.6915
9	0.1634	11.4590
10	0.2067	14.4992

The SEE for the KE with optimal bandwidths ranged from 0.0314 (at a score of 40) to 1.5153 (at a score of 4), following a typical shape of the KE SEE. The SEEs for the KE linear were U-shaped and were smaller than those from the KE with optimal bandwidths. They ranged from 0.1242 (at a score of 38) to 0.628 (at a score of 6).

Comparison with the operational equating results. Figure 1 plots the differences of converted scores between equipercentile equating and the KE with optimal bandwidths and the differences between linear equating and the KE linear. The difference is in raw score unit. It can be seen from the graph that the KE with optimal bandwidths produced very close results to its analogue equipercentile equating, except for some slight differences at the lower score range, but the observed differences were at the level of noise in the data as reflected by the SEE. The KE linear yielded almost identical results to linear equating.

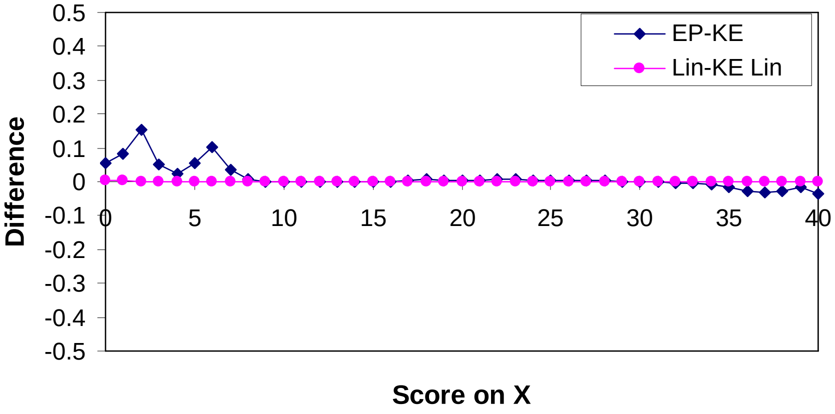

***Figure 1.* The differences of converted raw scores between KE and its target approximations in the EG design.**

Note. EP = equipercentile equating; Lin = linear equating; KE Lin = kernel equating linear method.

Table 10 gives the summary measure of differences between the KE and its target approximations for the EG design. Root mean squared difference (RMSD) is defined as Equation 6:

$$RMSD = \sqrt{\bar{d}^2 + sd_d^2}\ , \qquad (6)$$

where $\bar{d}$ is the mean of the differences and *sd* is the standard deviations of the differences. These summary statistics show that the differences between the KE and its target approximations were very small, especially in the linear case. All these indicated that the KE approximated well the operational equating methods in the EG design case for this particular data set.

Table 10

Summary Measure of Differences Between KE and Its Target Approximations in the EG Design

Summary	EP–KE	Lin.–KE Lin.
Mean diff.	0.0113	0.0004
SD diff.	0.0354	0.0007
Max diff.	0.1547	0.0022
Min diff.	-0.0370	-0.0002
RMSD	0.0371	0.0008

Note. EP = equipercentile equating; Lin = linear equating; KE Lin = kernel equating linear method; RMSD = root mean squared difference.

NEAT Design With Internal Anchor

KE results. In the PSE, we used the presmoothed bivariate distributions based on the (5 + 5 + 1) model and a weight *w* of 0.5 to estimate marginal distributions of $X + A$ and $Y + A$ on the target population *T*. The optimal continuization values were 0.7120 for h_x and 0.7075 for h_y. For the KE linear case, we used bandwidths of 10σ for both distributions ($h_x = 134.3$, $h_y = 135.7$).

Table 11 shows the PRE values for the KE optimal and the KE linear for equating *X* to *Y* using the poststratification method. The PRE values for the KE optimal indicated a good match between the equating function computed at the discrete values of *X* and the target distribution of *Y*. The PRE values for the KE linear indicated a relatively close match between the linear equating function at the discrete values of *X* and *Y*, too. This gave some evidence for the similar shape of the two distributions.

The SEE for the KE with optimal bandwidths ranged from 0.5259 (at a score of 74) to 152.7801 (at a score of 17) following a typical shape of the KE SEE. It is important to note that there were no data below the score of 38 on both tests. Frequencies that were used to calculate SEEs for this score range were an artifact of the smoothing and continuization processes and were based on the model and on data available from other portion of the score range. The SEEs for the KE linear were U-shaped and were smaller than those from the KE with optimal bandwidths. They ranged from 0.3906 (at a score of 73) to 1.9769 (at a score of 0).

Table 11

The PRE Values for the KE Optimal and KE Linear for Equating X to Y in the NEAT Design

Moments	PRE (optimal)	PRE (linear)
1	0.0000	0.0000
2	0.0000	0.0000
3	0.0002	0.0347
4	0.0007	0.1258
5	0.0015	0.2857
6	0.0027	0.5211
7	0.0042	0.8341
8	0.0062	1.2243
9	0.0085	1.6893
10	0.0112	2.2260

Comparison with the operational equating results. Figure 2 plots the differences of converted scores between the FE method and KE PSE and between the Tucker method and KE linear method in the NEAT design. Differences were in the raw score unit, and those greater than 0.5 were not plotted due to the range of the vertical axis (the largest difference is 1.1 at the score of 0 between FE and KE PSE). It can be seen from the graph that KE PSE with optimal bandwidths produced very close results to its analogue, FE, except at the lower score range. Actually, the observed frequencies for scores below 38 were zero on both forms. Therefore, the slight differences at the lower score range might not be of practical significance at all. Figure 2 also shows that the Tucker method and the KE linear method yielded very similar results. This indicated that the regression of tests on the anchor might be linear.

Table 12 gives the summary measure of differences between the KE and its target approximations in the NEAT design. It can be seen that the differences between the KE and its target approximations were very small, especially in the linear case. All these indicated the KE approximated well the operational PSE methods in the NEAT design with internal anchor.

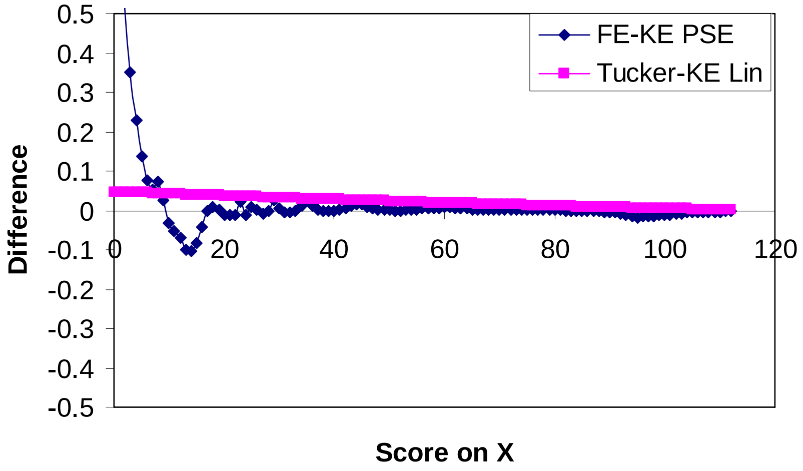

Figure 2. **The differences of converted raw scores between KE and its target approximations in the NEAT design with internal anchor.**

Note. FE = frequency estimation equipercentile equating; KE Lin = KE linear method; PSE = poststratification equating.

Table 12

Summary Measure of Differences Between KE and its Target Approximations in the NEAT Design With Internal Anchor

Summary	FE–KE PSE	Tucker–KE Lin
Mean diff.	0.0243	0.0219
SD diff.	0.1364	0.0135
Max diff.	1.0556	0.0472
Min diff.	–0.1048	0.0010
RMSD	0.1386	0.0257

Note. FE = frequency estimation equipercentile equating; Lin = Linear; PSE = poststratification equating; RMSD = root mean squared difference.

The operational equating methods used include the chained linear and chained equipercentile equating methods. Table B1 in Appendix B shows the equated scores corresponding to the selected raw scores for each operational equating method and KE method.

NEAT Design With External Anchor

KE results. In the PSE involving the NEAT design with external anchor, we used the presmoothed bivariate distributions based on the (3 + 3 + 1) model and a weight *w* of 0.5 to estimate marginal distributions of *X* and *Y* on the target population *T*. The optimal continuization values were 0.6686 for h_x and 0.6935 for h_y. For the KE linear case, we used bandwidths of 120 for both distributions (approximately 10σ).

Table 13 shows the PRE values for the KE optimal and KE linear for linking *X* to *Y* using the PSE method. The PRE values for the KE optimal indicated a good match between the equated function computed at the discrete values of *X* and the target distribution points of *Y*. The PRE values for the KE linear indicated a poor match between the linear equating functions at the discrete values of *X* and *Y*. This was a result of the different shapes of the two distributions.

Table 13

The PRE Values for the KE Optimal and KE Linear for Linking X to Y in the NEAT Design

Moments	PRE (optimal)	PRE (linear)
1	0.0003	0.0000
2	–0.0013	–0.0002
3	–0.0010	0.4694
4	0.0024	1.6724
5	0.0098	3.7749
6	0.0212	6.8941
7	0.0367	11.1235
8	0.0561	16.5478
9	0.0794	23.2517
10	0.1062	31.3248

The SEE for the KE with optimal bandwidths ranged from 1.1866 (at a score of 54) to 9.4639 (at a score of 24), following a typical shape of the KE SEE. The SEEs for KE linear are U-shaped, and they ranged from 1.0329 (at a score of 54) to 9.1628 (at a score of 10).

Comparison with the operational equating results. Figure 3 plots the differences of converted scores between FE and KE PSE and between the Tucker method and KE linear method in the NEAT design with external anchor. Differences were in the raw score unit, and those greater than 0.5 were not plotted due to the range of the vertical axis (the largest difference was 3.4 at the score of 0 between FE and KE PSE). It can be seen from the graph that KE PSE with optimal bandwidths produced similar results to its analogue, FE, over the range of ability at which most examinees scored. The Tucker method and the KE linear method produced very close results. This indicated that the regression of the tests on the anchor might be linear.

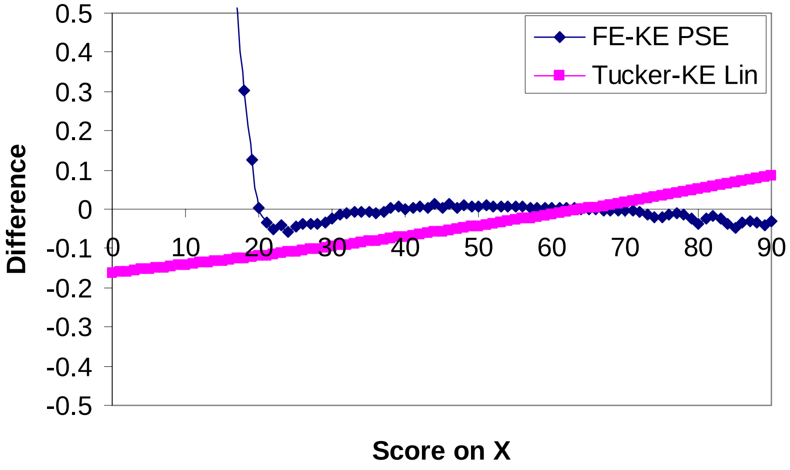

***Figure 3.* The differences of converted raw scores between KE and its target approximations in the NEAT design with external anchor.**

Note. FE = frequency estimation equipercentile equating; KE Lin = KE linear method; PSE = poststratification equating.

Table 14 gives the summary measure of differences between the KE and its target approximations for this NEAT design. The differences between the KE and its target approximations were very small in the linear case. In the nonlinear case, the RMSD was bigger than that in the other designs. This was mainly caused by the relatively large values of differences in the lower score range. Again, since no examinee scored below 20 on either form, these relatively large differences in the lower score range might not be of practical significance at all.

The operational equating methods used also include chained linear and chained equipercentile equating methods. Table B2 in Appendix B shows the equated scores corresponding to the selected raw scores for each operational equating method and related KE method.

Table 14

Summary Measure of Differences Between KE and its Target Approximations in the NEAT Design With External Anchor

Summary	FE–KE PSE	Tucker–KE Lin
Mean diff.	0.3972	–0.0505
SD diff.	0.8990	0.0725
Max diff.	3.4198	0.0852
Min diff.	–0.0571	–0.1624
RMSD	0.9829	0.0883

Note. FE = frequency estimation equipercentile equating; KE Lin = KE linear method; PSE = poststratification equating; RMSD = root mean squared difference.

It should be pointed out that if the requirements for equating are not well met, employing a sophisticated, almost automatic statistical methodology such as the KE method will not circumvent the problem. In this linking case, the item format of the external anchor was different from that of the test, and the correlations between the test and the anchor were relatively low. Therefore, one should be careful in interpreting the results in this linking case.

Discussion and Conclusion

This study applied KE in three equating designs with various sample sizes and characteristics and compared the KE results with the results from the operational equating methods that the KE approximates. The results in general indicate that KE can approximate the traditional equating methods well. In the EG design, when optimal bandwidths are used, the KE results are very close to those of the traditional equipercentile equating methods. When large bandwidths are used, KE yields almost identical results to the traditional linear equating method for this data set. In the NEAT design with internal anchor, the KE version of PSE methods with optimal bandwidths produces very close results to its analogue, FE, in most of the score range. The KE with large bandwidth produces even closer results to the Tucker method. Similar results are found in the NEAT design with external anchor, except that at lower score range, the differences between the results of the KE with optimal bandwidths and those of the FE method are relatively large, but still at the noise level.

The fact that the KE with large bandwidth produces almost identical results to the linear method in the EG design is within expectation. It has been mentioned previously that when conducting KE, the score distributions are presmoothed with the same log-linear models as those used in the operational equatings. Linear equating is based on the means and variances of the presmoothed score distributions. The KE function with large bandwidths is approximately linear, in which case the estimated KE function is mainly based on the means and standard deviations of the continuized distributions. Because the mean and variances of the continuized distribution in the KE exactly match those of the presmoothed discrete distribution (von Davier et al., 2004b), it is not surprising that the KE function with large bandwidth is very close to the linear equating function in the EG design. For similar reasons, the KE with large bandwidth should produce very similar results to its analogue, the PSE-linear method, in the NEAT design using the PSE method. The fact that it also produces close results to the Tucker method in this study indicates that the assumptions for the Tucker method may be satisfied (i.e., the regression of *X* and *Y* on *V* are approximately linear, and the conditional variances are almost constant).

For all three equating designs, the largest differences between the KE with optimal bandwidths and its analogues in the operational equating methods occurred at the lower score range, where few examinees in the sample scored. These differences may result from the different approaches of the two methods to estimate the equating relationships at the lower score range. The continuization process in KE may spread the continuous distribution of probability beyond the range of the raw scores, whereas the traditional equipercentile equating method does not. For the NEAT design with an external anchor, the test forms have low reliabilities, and the correlations of the tests with the anchor are low as well. In this case, any equating results are questionable, especially those equating methods that use the anchor as a covariate (the PSE methods). The KE method will not circumvent this problem, either. Therefore, the results in this case should be interpreted carefully. Livingston's (1993) study showed that KE produces slightly more accurate results than chained equipercentile equating methods. Because the true equating relationships are unknown in this study, we do not know whether the operational equating methods or KE produced more accurate results at the lower score range. However, at these score points the data are sparse and the noise is large; therefore, the relatively large differences at these score points may not be of practical significance in these equating situations.

By using operational equating data from various designs, this study shows that the KE can approximate the results of the traditional equating methods well. The results from this study, together with the results from the other evaluation study that used pseudo-tests (von Davier et al., 2005) and the results from Livingston’s (1993) resampling study, show the promise of the KE method and increase the confidence of the practitioners in this method.

Despite this, more research needs to be done to evaluate the performance of the KE in equating situations different from those studied here. Given the results of this study, we consider that another aspect that needs more research is the accuracy of the results produced by the KE at score points where few examinees scored.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: ETS. (Originally published 1971)

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.

Brent, R.P. (1973). *Algorithms for minimization without derivatives.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

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 chain and post-stratification 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), 15– 32.

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

von Davier, A. A., Holland, P. W., Livingston, S., Casabianca, J., Grant, M., & Martin, K. (2005, April). *An evaluation of the kernel equating method: A special study with pseudo-tests from real test data*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

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

ETS. (2004b). KE-Software [computer software]. Princeton, NJ: Author.

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions (ETS RR-87-31). Princeton NJ: ETS.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS RR-89-7). 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: Methods and practices*. New York: Springer.

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

Appendix A

An Outline of KE

Here we give a brief outline of the KE method of observed-score test equating (von Davier et al., 2004a). The KE method is discussed in detail in von Davier et al. (2004b) for all of the standard equating designs. As mentioned before, the KE method has five basic steps:

Step 1: 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 log-linear models for univariate and bivariate score distributions, as discussed in Holland and Thayer (2000), because of their 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; 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 log-linear 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, there are formulas (see Equations A1 and A2), depending on the equating design, that use the smoothed score distribution estimates to produce estimates of the score probability distributions on *T*, which we call *r* and *s*, where

$$r_j = P\{X = x_j|T\},\ s_k = P\{Y = y_k|T\} \qquad \text{(A1)}$$

and the vectors *r* and *s* are given by

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

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 (von Davier et al., 2004a, chapter 2).

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 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. The KE approach is to make this step explicit and to make the sense of the approximation clear. Older 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 in Equation A3:

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

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

$\Phi(z)$ 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 $\{x_j\}$. μ_{XT} and σ^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 new and continuized 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 discrete score probabilities (that have been previously estimated by the log-linear models).

We have found the following penalty functions (see Equation A4) to give good results.

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

where d_j is the width of the interval associated with the score x_j (often these widths are all set equal to 1). We also place a penalty of 1 for every score point where the density $f_T(x;\, h)$ is U-shaped around it, as shown in Equation A5:

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

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. What "near" means is a parameter of PENALTY$_2(h)$, and we can combine the two penalties with a weight, as shown in Equation A6:

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

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. Brent's algorithm (Brent, 1973) 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 shown in Equation A7:

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

where $G_T^{-1}(p;\, h_Y)$ denotes the inverse of $p = G_T(y;\, h_Y)$.

The function $e_{XY}(x)$ 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 the moments of the two distributions, $e_{XY}(X)$ and Y (up to the 10th), via the PRE in the p$^{\text{th}}$ moment formula (see Equation A8):

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

$$\text{PRE}(p) = 100\text{x}[\mu_p(\text{e}_\text{Y}(X)) - \mu_p(Y)]/\ \mu_p(Y). \qquad \text{(A8)}$$

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

The SEE for $e_{XY}(x)$ depends on three factors that correspond to the above four steps—(a) presmoothing, (b) computing r and s from the smoothed data, and (c) 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. 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 SEE difference discussed in von Davier et al. (2004b) is a new tool, unique to KE, for evaluating the degree to which KE and linear equating agree. Moreover, the SEE difference can be used to assess the difference between two equating functions that depend on the same parameters.

Appendix B

Equated Scores Corresponding to Selected Raw Scores, NEAT Design

Table B1

Equated Scores Corresponding to Selected Raw Scores, by Each Equating Method, NEAT Design With Internal Anchor

Raw score on Form *X*	43	63	73	82	102
(percentage below on X in P)	< 1%	23.5%	50%	73%	99%
Corresponding score on Form *Y*, as determined by					
Chained linear	51.16	71.47	81.62	90.76	111.07
Tucker	51.21	71.50	81.64	90.77	111.06
Kernel, poststratification, large bandwidth	51.18	71.48	81.63	90.76	111.05
Levine observed-score	51.09	71.43	81.60	90.75	111.08
Chained equipercentile	50.94	71.42	81.67	90.92	111.25
Frequency estimation equipercentile	50.37	71.60	81.86	90.86	110.23
Kernel, poststratification, optimal bandwidth	50.36	71.60	81.86	90.86	110.24

Table B2

Equated Scores Corresponding to Selected Raw Scores, by Each Equating Method, NEAT Design With External Anchor

Raw score on Form *X*	39	50	56	64	82
(percentage below on X in P)	0%	27%	49%	73%	99%
Corresponding score on Form *Y*, as determined by .					
Chained linear	36.90	50.07	57.26	66.84	88.40
Tucker	36.15	49.51	56.80	66.51	88.37
Kernel, poststratification, large bandwidth	36.23	49.55	56.82	66.51	88.31
Chained equipercentile	35.17	50.71	58.21	67.31	83.52
Frequency estimation equipercentile	33.37	50.21	58.01	67.10	83.09
Kernel, poststratification, optimal bandwidth	33.36	50.20	58.01	67.10	83.11