Research Report

An Exploration of Kernel Equating Using SAT® Data: Equating to a Similar Population and to a Distant Population

Jinghua Liu

Albert C. Low

Research & Development

April 2007
RR-07-17

An Exploration of Kernel Equating Using SAT® Data: Equating to a Similar Population and to a Distant Population

Jinghua Liu and Albert C. Low
ETS, Princeton, NJ

April 2007

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

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

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

Abstract

This study applied kernel equating (KE) in two scenarios: equating to a very similar population and equating to a very different population, referred to as a distant population, using SAT® data. The KE results were compared to the results obtained from analogous classical equating methods in both scenarios. The results indicate that KE results are comparable to the results of other methods. Further, the results show that when the two populations taking the two tests are similar on the anchor score distributions, different equating methods yield the same or very similar results, even though they have different assumptions.

Key words: Kernel equating (KE), SAT using KE, KE and classical equating comparison

Acknowledgments

The authors thank Ning Han, Miriam F. Cahn, and Annie Nellikunnel for their help with equatings.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

For decades test scores have been equated to produce matched score distributions so that scores can be comparable across different test forms. Researchers in the field have devoted much energy toward developing equating methods. Holland and Thayer (1989) and von Davier, Holland, and Thayer (2004a) developed a new approach to observed-score equating, kernel equating (KE). A distinctive difference between KE and traditional equipercentile equating is that score distributions are converted from discrete to continuous distributions in KE by use of Gaussian kernel smoothing, as opposed to the linear interpolation used in the traditional percentile rank approach. "KE is a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contains the linear equating as a special case" (von Davier et al., 2004a, p. 45). Hence, KE holds the promise of approximating other equating methods while providing new measures of statistical accuracy.

At the 2005 annual meeting of National Council on Measurement in Education (NCME), a symposium titled *The Kernel Method of Test Equating: Evaluation and Applications* explored KE from different perspectives using either real data from different testing programs or simulated data. One of the topics discussed that is related to the present study was: How closely does KE approximate other equating methods? KE results were shown to be comparable to those obtained using traditional equating methods (von Davier et al., 2005).

This study examines the results of KE versus those of other equating methods in two different scenarios: equating to a very similar population and equating to a quite different population. It compares the results from different equating methods based on different assumptions (e.g., chained equating and post-stratification equating) across the two scenarios using SAT® data for the purpose of illustration. The second section of the study reviews relevant theoretical background and previous studies. The third section describes the SAT equating procedure and pairwise comparisons between KE and its analogues. The fourth section presents the results of the study, and the fifth section synthesizes the findings.

Theoretical Background and Previous Research

This section first reviews the differences between two types of equating methods (chained equating and post-stratification equating) based on different assumptions under a non-equivalent-groups anchor test (NEAT) design. Then the report briefly summarizes the five steps

of KE. Each KE method and its analogue in the classical equating family are compared and previous studies are reviewed briefly.

Chained Equating Versus Post-Stratification Equating

In the NEAT design, two different populations, *P* and *Q*, take two different test forms, *X* and *Y*, respectively, and an anchor test *V* is used to link them. von Davier, Holland, and Thayer (2004a, 2004b) discussed differences between the two types of equating methods used for the NEAT design: chained equating and post-stratification equating. Chained equating uses the anchor as part of a chain: first link *X* to *V* on *P*, and then link *V* to *Y* on *Q*. The two linking functions are then composed to map *X* to *Y* through *V*. There are two chained equating methods: the chained linear method and the chained equipercentile method.

Chained linear equating assumes that the mean/sigma linking relationship between *X* and *V* would be the same if it were observed on *Q*. Likewise, it assumes that the mean/sigma linking relationship between *Y* and *V* would be the same if it were observed on *P*. The chained equipercentile method first converts scores *X* to *V* on *P* using the equipercentile method, resulting in the equipercentile function $e_{VP}(x)$. It then finds the equipercentile equating relationship for converting *V* to *Y* on *Q* and gets the resulting function $e_{YQ}(v)$. To equate *X* to *Y*, it is assumed that the equipercentile equating relationship between *X* and *V* (or between *Y* and *V*) would be the same if it were observed on *Q* (or on *P*), and the method converts *X* to *V* using $e_{VP}(x)$. It then equates the resulting score *V* to *Y* using $e_{YQ}(v)$ (Kolen & Brennan, 2004).

Post-stratification equating uses the anchor test *V* to estimate the distribution of *X* on *Q* and the distribution of *Y* on *P*. It assumes that the conditional distribution of *X* given *V* and the conditional distribution of *Y* given *V* are population invariant and then post-stratifies the distributions of both *X* and *Y* on a target population *T* (synthetic population of *P* and *Q*). The post-stratification equating methods used in this study are the Tucker (linear) method and frequency estimation (nonlinear) method.

The Tucker method assumes that the regression of *X* on *V* is linear and population invariant. The conditional variance of *X* given *V* is also assumed to be the same for *P* and *Q*. A similar set of assumptions is made about the regression of *Y* on *V* and about the conditional variance of *Y* given *V*. The frequency estimation method requires conditional score distributions and assumes that, for both *X* and *Y*, the conditional distribution of the total score, given each

anchor score, $V = v$, is population invariant. The more similar the two populations, the more likely this assumption will hold (Kolen & Brennan, 2004).

Kernel Equating

The five basic steps of KE are summarized below (von Davier et al., 2004a).

Step 1: Presmoothing. First, data are presmoothed using various techniques. In the NEAT design, estimates of bivariate score distributions and the C-matrices are obtained. The latter are covariance matrices of the estimated distributions used to compute the standard error of equating and the standard error of equating difference.

Step 2: Estimation of the score probabilities. Once data are smoothed, the estimates of score probabilities are produced on the target population through a design function. For chained equating in the NEAT design, two design functions from P and Q are combined into a single function. For post-stratification equating in the NEAT design, the marginal distributions of *X* and *Y* on the target population *T* are first estimated, and then the design function is computed from the two marginal distributions.

Step 3: Continuization. This step involves converting discrete cumulative distribution functions (CDFs) of *X* and *Y* into approximating continuous CDFs. As mentioned in the introduction, a distinctive feature of KE is that while the traditional percentile rank approach to equipercentile equating uses linear interpolation KE uses Gaussian kernel continuization to estimate a continuous CDF of *X*, $F_{hX}(x)$ and a continuous CDF of *Y*, $G_{hY}(y)$. Gaussian kernel continuization involves a choice of bandwidths, h_X and h_Y. von Davier et al. (2004a) recommended using a penalty function to select the bandwidths to make the density functions smooth and to closely fit the discrete cumulative functions. The size of the bandwidth is directly related to the curvilinearity/linearity in the resulting equating function. A curvilinear function is computed using optimal bandwidths, which are chosen to closely fit the discrete cumulative function. A linear equating function is computed using bandwidths that are much larger than the optimal bandwidths (e.g., $h_X > 10\sigma_X$ and $h_Y > 10\sigma_Y$).

Step 4: Equating. The *X*-to-*Y* equipercentile equating function can be computed from the continuized CDFs, $F_{hX}(x)$ and $G_{hY}(y)$, using the formula: $e_Y(x) = G_{hY}^{-1}(F_{hX}(x))$. The selection of bandwidth determines the linear or curvilinear equating functions.

Step 5: Calculating standard error of equating. The standard error of equating is estimated based on the delta method. The standard error of equating depends on the fit of the presmoothing, the estimation of score probabilities, and the equating function. In addition, KE provides the standard error of equating difference to evaluate the differences between two equating functions.

Pairwise Comparisons of Kernel Equating and Its Analogue

For each of the classical equating methods, KE provides a version that approximates the classical analogue. This section looks at pairwise comparisons to be made between the particular KE results and its analogue based on one of the traditional anchor equating methods.

Comparison 1: Chained equating versus kernel version of chained equating. The kernel version of chained equating with a large bandwidth approximates chained linear equating. The kernel version of chained equating with optimal bandwidth approximates chained equipercentile equating.

Comparison 2: Post-stratification equating versus kernel version of post-stratification equating. Post-stratification equating methods include the Braun-Holland linear method, the Tucker linear method, and the frequency estimation method. (The Braun-Holland method and Tucker method are identical when the regressions on *V* are linear and the conditional variances are constant. In this study only the Tucker method and the frequency estimation method were tried.) The kernel version of post-stratification equating with large bandwidth approximates the Tucker linear method, given that the Tucker assumption about the linearity of the regression holds. The kernel version of post-stratification equating with optimal bandwidth approximates the frequency estimation method.

Previous Studies

Several studies have been conducted to evaluate KE results by comparing them to results produced by traditional equating methods. Using real data from *The Praxis Series: Professional Assessments for Beginning Teachers*®, Mao et al. (2005) compared KE results to those of other equating methods. Two equating designs were employed: the equivalent groups (EG) design and the non-equivalent groups anchor test (NEAT) design. In the EG design, new form *X* was equated to old form *Y* through equivalent groups. KE, equipercentile equating, and linear equating were applied. The results showed that the differences between KE with optimal

bandwidth (KE nonlinear) and equipercentile equating were very small, as were the differences between KE with large bandwidth (KE linear) and linear equating. In the NEAT design, new form *X* was equated to old Form *Y* through both an internal and an external anchor. KE, frequency estimation, and Tucker methods were used. The differences between the KE version of post-stratification equating with optimal bandwidth (KE PSE nonlinear) and the frequency estimation method were trivial. The differences between the KE version of post-stratification equating with large bandwidth (KE PSE linear) and Tucker equating were also small, in both the internal anchor equating case and the external anchor equating case.

The results from Mao et al. (2005) indicate that KE can approximate traditional equating methods well. However, some issues need further investigation:

1. In the Mao et al. (2005) study, the new form equating sample and the old form equating sample in the NEAT design were very similar in ability level. The standardized mean difference on the anchor test was .01 on the internal anchor and .11 on the external anchor. When an anchor test is used for equating, the anchor is a miniature of the total test form, and the groups taking the old and new form are similar to each other, all equating methods tend to agree with each other and give similar results. On the other hand, when an anchor test is used and the groups taking the old and new forms are quite different, any equating method may give poor results (Kolen, 1990). Using SAT data, Dorans, Liu, and Hammond (2004) found that equating results from different methods are tightly bunched when there are small ability differences between groups and diverge when there are large ability differences. Hence, the question is: When there is a large ability difference between the new form group and the old form group, will KE still approximate the results of traditional equating methods? How do KE results look across different equating methods (e.g., post-stratification vs. chained equating, nonlinear vs. linear equating)?

2. The post-stratification methods, including the KE versions of post-stratification, frequency estimation, and Tucker equating, were examined in the current study. The post-stratification methods work well when the stratification variable is highly related to the total test. In the Mao et al. study, the correlation between the anchor and the total test in the external anchor design was low, which could have contaminated the equating results, as the authors pointed out.

3. None of the chained equating methods was explored in the Mao et al. (2005) study.

Some of these issues were explored in another KE study (von Davier et al., 2005), which intended to answer two questions about KE: (a) How closely do the KE results approximate the results of other equating methods, and (b) are the KE results close to an equating criterion when compared to other methods? Two equating samples were drawn from two different administrations of an existing test, two pseudo tests and an external anchor were constructed, and the item responses from the real test were used. The results from KE were then compared to those derived from its traditional equating counterpart. In addition, the results of KE and other equating methods were compared to the equating criterion (the equipercentile equating in the combined samples) used in the study. The results showed that KE can approximate other methods very closely. Further, the KE results were closer to the equating criterion than were the results produced by other equating methods.

A major problem with the von Davier et al. study (2005) was that the two pseudo-tests were not parallel to each other. Even though they were very similar from a content perspective, one form was much more difficult than the other. The correlations between the anchor and the total tests were also low (.782 in the new group and .735 in the old group).

Like the Mao et al. (2005) and von Davier et al. (2005) studies, this study examines KE results versus traditional equating results. The features that distinguish this research from the previous studies are the inclusion and comparison of two equating scenarios: equating to a similar population where the new form group and the old form group were very close in ability level and equating to a very different population where the new form group and the old form group were quite different in ability. The correlation between the anchor and the total test was much higher than in the previous research. In addition, the results from different equating methods based on different assumptions (e.g., chained equating and post-stratification equating) were compared across the two different scenarios in this study.

Methodology

Equating Design and Equating Samples

In this study, we equated a new SAT verbal test to an old SAT verbal test. Both were built according to the same strict content and statistical specifications, and the two tests were parallel to each other. In a typical SAT administration, the new form is equated to multiple old SAT forms

though a NEAT design. This design has produced stable equating results because it acknowledges the important role that equating of the old form plays in placing a new form on scale.

In this study we chose two old forms. One was administered at the same time of year as was the new form it was compared to. The resulting equating is thus referred as being to a **s**imilar population. The other old form was administered at a different time of year, in this case one of the three core administrations of the SAT that contribute large numbers of scores to the SAT cohort. The resulting equating is said to be to a distant population. In the latter case, ability level differences between the two groups were usually large.

For the purposes of this study, KE was conducted using the KE stand-alone software (ETS, 2006); other equating methods were performed using GENASYS (ETS, 2006). Log-linear smoothing models (Holland & Thayer, 1987, 2000) were used. The model of (6, 6, 1) was determined to be the best fit. Two C-matrices, C_P and C_Q, were obtained for each of the two equatings.

Some Indices Used in This Study

Percent relative error. KE provides a tool, percent relative error (PRE), to assess how well an equating function, $e_Y(x)$, matches the discrete target distribution *Y*. KE compares up to the 10th moment of the two distributions via the PRE in the p^{th} moment.

Standard error of equating difference. KE also provides a tool to evaluate the differences between two equating functions: the standard error of equating difference (SEED). If two equating functions differ by more than twice the SEED over important ranges of the raw scores of *X*, this result can be regarded as evidence that the differences are significantly different from zero. If, on the other hand, the differences do not differ by more than twice the SEED, then they are not considered to be big enough to cause concern (e.g., the differences were caused by sampling errors).

Results

In this section, the following comparisons are made based on the results of equating to a **s**imilar population:

- the differences between kernel chained equating and traditional chained equating,
- the differences between kernel post-stratification equating and traditional post-stratification equating, and

- the differences between kernel chained equating and kernel post-stratification equating.

A similar set of comparisons is then evaluated on the equating to a distant population.

Equating the Verbal Test to a Similar Population

Table 1 presents the raw score summary statistics in the equating of form *X* to form *Y1* through the anchor *V1* to a similar population. These include new form group performance on test *X* and anchor *V1*, old form group performance on test *Y1* and anchor *V1*, the correlation between the anchor and the total test in each group, the reliability of the total test and the anchor, and so on. As the results show, the standardized mean difference on the anchor between the new and old form groups was nearly zero and the ratio of the variances was close to one, which indicates the two populations had the same score distributions on the anchor test. The correlation between the anchor and the total test was approximately .88 for both the new and the old forms, and the reliability was above .90 for both forms.

Table 1

Summary Statistics on the New-Form Sample and on the Old Form Sample in the Equating to a Similar Population

Statistics	New form sample		Old form sample	
	Test *X*	Anchor *V1*	Test *Y1*	Anchor *V1*
Sample size	7,528	7,528	6,837	6,837
Number of items	78	35	35	78
Mean	34.96	16.06	34.01	16.09
SD	17.82	7.82	17.19	7.65
Skewness	.13	.01	.23	.04
Kurtosis	2.22	2.41	2.43	2.51
Reliability	.93	.84	.92	.83
Correlation	.88		.87	
Standardized mean difference (new-old)			0.00	
Ratio of variances (new-old)			1.04	

Table 2 provides PRE values for the kernel version of chained equating and post-stratification equating, with optimal bandwidth (PRE optimal) and with large bandwidth (PRE linear). The optimal continuization values were 0.7390 for h_X and 0.7610 for h_Y. The linear version of KE used bandwidths of 200 (approximately $10\,\sigma_X$ and $10\,\sigma_Y$). The PRE values for the KE linear version were larger than those for the KE optimal, except for the first two moments.

Table 2

PRE Values for Kernel Equating Optimal Versus Linear in Equating to a Similar Population

Moments	Chained equating				Post-stratification equating	
	PRE optimal		PRE linear		PRE optimal	PRE linear
	X to *V1*	*V1* to *Y1*	*X* to *V1*	*V1* to *Y1*		
1	–0.0001	0.0021	0.0000	0.0000	0.0011	0.0000
2	–0.0034	0.0074	–0.0004	0.0000	0.0004	0.0000
3	–0.0044	0.0113	0.7868	–1.3776	–0.0079	–0.6250
4	0.0052	0.0075	1.6783	–3.3504	–0.0171	–1.9962
5	0.0261	–0.0049	2.3918	–5.5634	–0.0238	–3.9651
6	0.0597	–0.0273	2.8006	–7.7846	–0.0255	–6.3487
7	0.1074	–0.0604	2.8749	–9.9044	–0.0207	–8.9880
8	0.1702	–0.1046	2.6315	–11.8770	–0.0081	–11.7628
9	0.2490	–0.1603	2.1086	–13.6927	0.0128	–14.5880
10	0.3446	–0.2275	1.3510	–15.3586	0.0426	–17.4055

Note. PRE = Percent relative error.

Comparison of kernel version of chained equating to traditional chained equating. Figure 1 displays the differences between the raw-to-raw equatings of chained linear and KE CE linear, and also the differences between chained equipercentile and KE CE optimal. As can be seen from the graph, KE linear produced almost identical results to its analogue, chained linear, with a difference line of zero. KE optimal produced very similar results to chained equipercentile, except below Score Level 1. There the differences started to deviate from zero, with a maximum difference of approximately 0.2 (except at Score –19).

Table 3 presents the summary statistics of raw-to-raw conversions based on KE methods and their target analogues. The results for chained equating show that the differences between

both KE linear and chained linear and between KE optimal and chained equipercentile are so small as to be nearly identical.

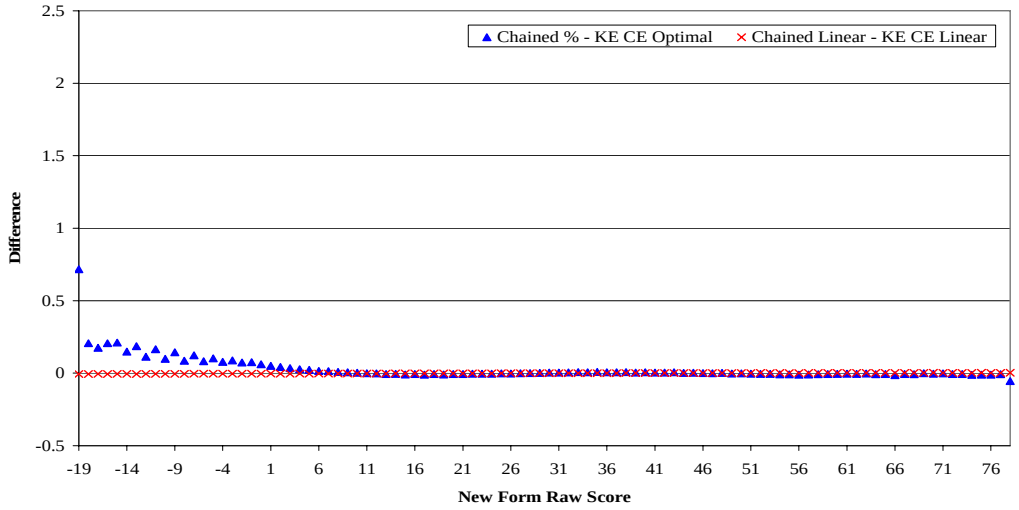

***Figure 1*. Equating to a similar population: raw-to-raw equating between kernel equating and its target approximates in chained equating.**

Table 3

Summary Statistics of Raw-to-Raw Conversions of Kernel Equating and Its Analogues in the Equating to a Similar Population

Statistics	Chained equating				Post-stratification equating			
	KE linear	Chained linear	KE optimal	Chained %	KE linear	Tucker	KE optimal	Freq. est.
Mean	33.94	33.94	33.99	33.99	34.00	33.95	34.01	34.01
SD	17.56	17.56	17.53	17.52	17.44	17.47	17.44	17.44
Min.	–19.22	–19.22	–20.18	–19.47	–18.81	–18.95	–21.64	–19.49
Max.	76.35	76.36	78.16	78.11	76.13	76.15	78.43	78.12

Note. KE = kernel equating, freq. est. = frequency estimation.

Comparison of kernel version of post-stratification equating to traditional post-stratification equating. Figure 2 displays the raw-to-raw equating differences between Tucker and KE PSE linear and between frequency estimation and KE PSE optimal. The results show that KE PSE linear produced very similar results to Tucker. KE PSE optimal also produced results close to those from frequency estimation, except below Score Level 1, where the differences increased.

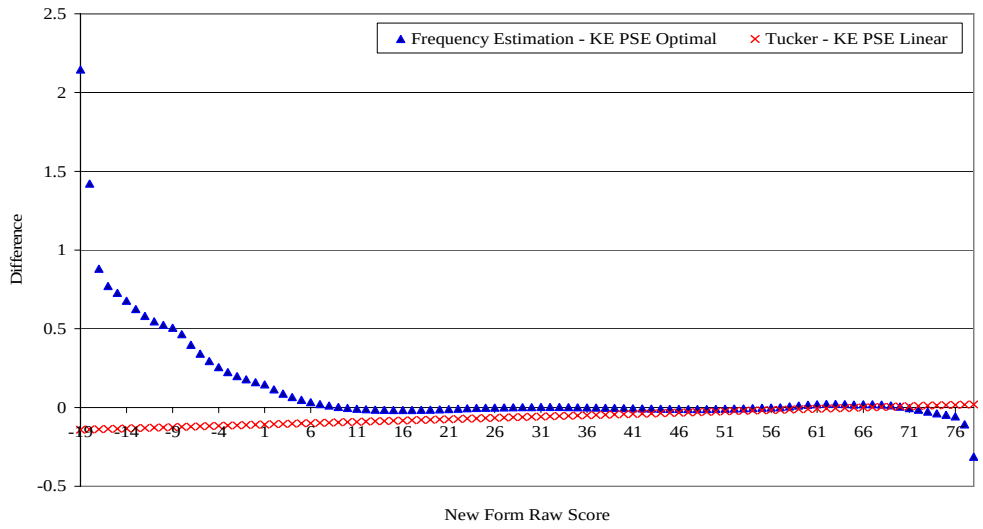

***Figure 2*. Equating to a similar population: raw-to-raw equating differences between kernel equating and its target approximates in post-stratification equating.**

Table 3 also contains summary data for the post-stratification equating comparisons between KE and its analogues. The results show that the differences between KE PSE linear and Tucker and between KE PSE optimal and frequency estimation were very small.

Overall, each pairwise comparison between a KE method and its analogue, whether chained or post-stratification, linear or optimal, is a good match. Given the small standard error of equating (SEE) for each KE method (less than 0.5 across most of the scale range), the observed differences are most likely noise.

Comparison of KE functions for chained equating and post-stratification equating. Figure 3 displays the SEE for both KE CE linear and KE PSE linear as nearly identical U-shaped curves ranging from approximately 0.15 to 0.45. Figure 4 plots the SEED between CE linear and PSE linear. Below Score Level 45, the CE linear conversion is lower than the PSE linear conversion, suggesting that CE assesses new form *X* as being an easier test than PSE does. Above Score Level 45, CE linear equates higher converted scores than PSE linear, indicating that CE measures new form *X* as being harder than PSE does. Regardless, the differences are within the range of –0.4 to 0.2, and the difference line lies within ±2 SEED across the entire score range. This suggests that the differences between the KE CE and KE PSE linear functions are from sampling errors and are not significantly different from zero.

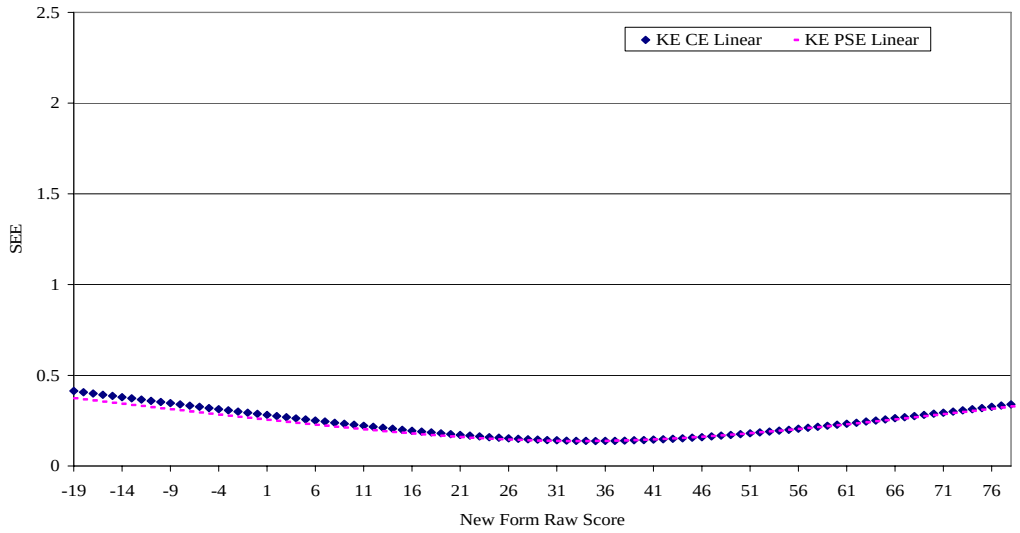

Figure 3. **SEE for equating to a similar population: KE CE linear versus KE PSE linear.**

Note. SEE = standard error of equating, KE = kernel equating, CE = chained equating, PSE = post-stratification equating.

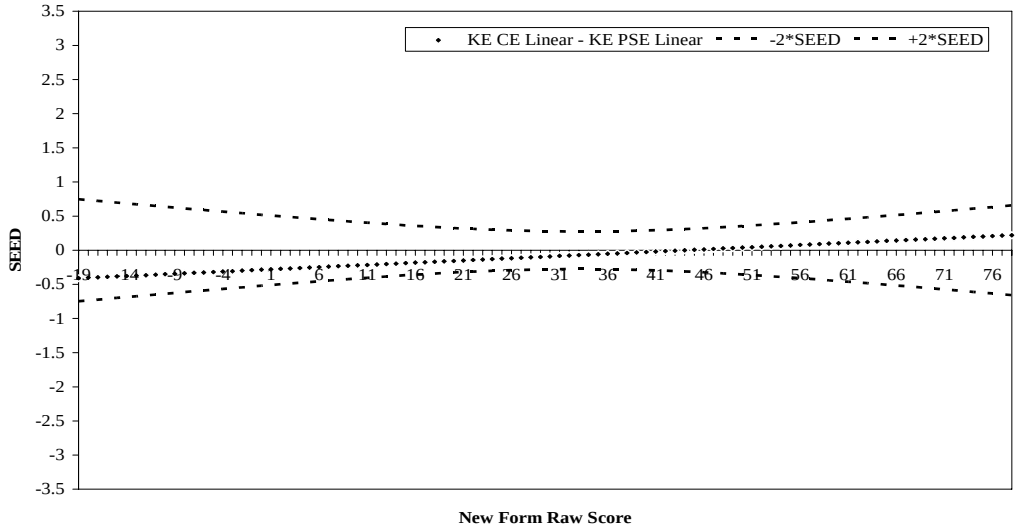

Figure 4. **SEED for equating to a similar population: KE CE linear versus KE PSE linear.**

Note. SEED = standard error of equating difference, KE = kernel equating, CE = chained equating, PSE = post-stratification equating.

Figures 5 and 6 show the differences between the nonlinear functions of KE CE and KE PSE. Again, the SEE values are very close to each other, as shown in Figure 5, and smaller than 0.5 across most of the score range. Figure 6 shows that the SEED intertwines with the zero line for

most of the *X*-values. Even though the SEED starts to increase below Score Level 1, the line is still within the ± 2 SEED band. Again, this nonlinear comparison shows that the differences between KE CE and KE PSE are due to sampling errors and are not significantly different from zero.

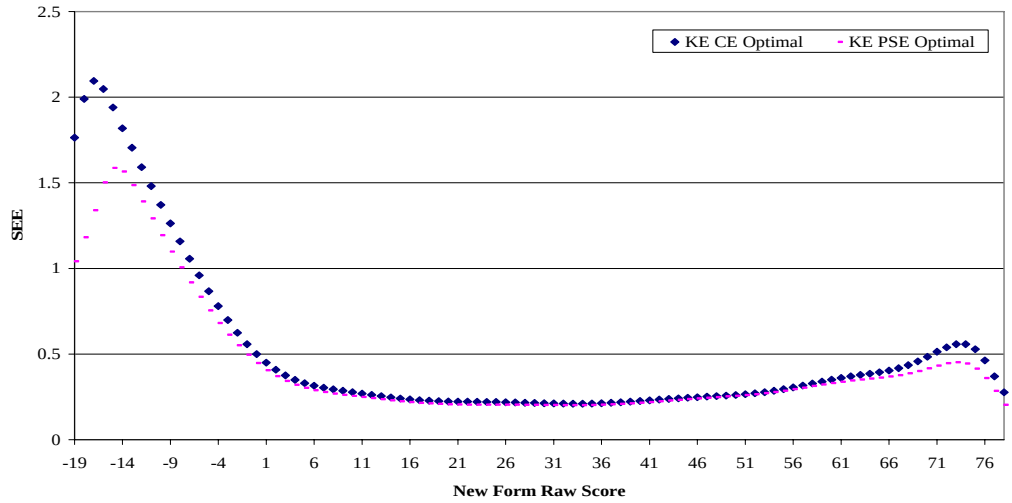

Figure 5. **SEE for equating to a similar population KE CE optimal versus KE PSE optimal.**

Note. SEE = standard error of equating, KE = kernel equating, CE = chained equating, PSE = post-stratification equating.

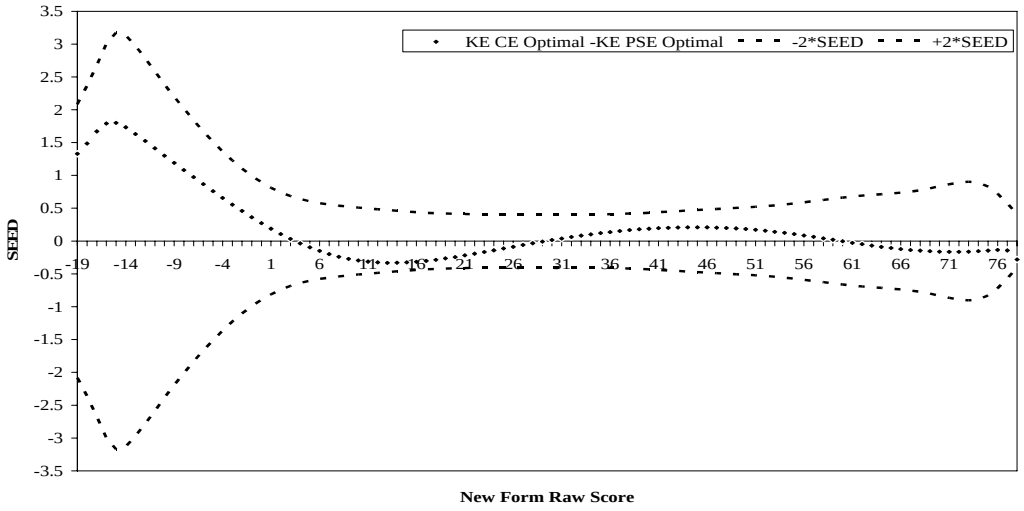

Figure 6. **SEED for equating to a similar population: KE CE optimal versus KE PSE optimal.**

Note. SEED = standard error of equating difference, KE = kernel equating, CE = chained equating, PSE = post-stratification equating.

The comparison suggests even though CE and PSE have different assumptions, it is possible the two methods produce identical or very similar results in certain circumstances. von Davier et al. (2004b) proved theoretically that: (a) when *P* and *Q* have the same score distributions on the anchor test, CE and PSE give the same results; and (b) when *X* and *Y* are perfectly correlated with the anchor, CE and PSE give identical results. In this case, since the correlations are 0.88 between *X* and the anchor and 0.87 between *Y* and the anchor, which are not perfect, the nearly identical results between CE and PSE can be judged to be due to the similarity between *P* and *Q* on the anchor. The standardized ability difference on the anchor is zero (see Table 1).

Equating the Verbal Test to a Distant Population

Table 4 presents the raw score summary statistics for the same verbal form *X* equated to a different old form (*Y2)* to a distant population through a different anchor (*V2).* As can be seen, the new form group was much less able than the old form group, with anchor means of 15.78 and 17.97, respectively. The standardized mean difference on the anchor was –.26. The correlation between the anchor and the total test was .87 and .89 for *X* and *Y2*, respectively. The reliability was above .90 for both total tests and around .85 for the anchor.

Table 4

Summary Statistics on the New Form Sample and on the Old Form Sample in the Equating to a Distant Population

Statistics	New form sample		Old form sample	
	Test *X*	Anchor *V2*	Test *Y2*	Anchor *V2*
Sample size	6,974	6,974	11,361	11,361
No. of items	78	35	35	78
Mean	34.78	15.78	39.78	17.97
SD	17.49	8.45	17.21	8.29
Skewness	.12	.08	–.13	–.10
Kurtosis	2.26	2.21	2.26	2.24
Reliability	.91	.83	.93	.86
Correlation	.87		.89	
Standardized mean difference (new-old)			–.26	
Ratio of variances (new-old)			1.04	

Table 5 provides the PRE values for KE equatings. The optimal continuization values were 0.7828 for h_X and 0.7506 for h_Y. For the KE linear version, bandwidths of 200 were used. Again, the PRE values for the KE linear version were larger than those for the KE optimal, except for the first two moments.

Table 5

Post-Stratification Equating Values for Kernel Equating Optimal Versus Linear in the Equating to a Distant Population

Moments	Chained equating				Post-stratification equating	
	PRE optimal		PRE linear		PRE optimal	PRE linear
	X to *V2*	*V2* to *Y2*	*X* to *V2*	*V2* to *Y2*		
1	0.0005	–0.0032	0.00002	0.00002	–0.0004	0.0000
2	–0.0004	–0.0080	–0.00035	0.00006	–0.0013	–0.0001
3	–0.0025	–0.0196	0.32173	0.17218	–0.0004	0.4701
4	0.0121	–0.0488	0.96850	0.45632	0.0022	1.3646
5	0.0458	–0.1005	1.92066	0.78769	0.0065	2.6231
6	0.1001	–0.1780	3.11869	1.12250	0.0121	4.1816
7	0.1754	–0.2838	4.50736	1.43270	0.0189	5.9837
8	0.2721	–0.4196	6.04333	1.69987	0.0263	7.9828
9	0.3903	–0.5872	7.69682	1.91160	0.0341	10.1411
10	0.5305	–0.7879	9.44936	2.05901	0.0419	12.4292

Note. PRE = Percent relative error.

Comparison of kernel version of chained equating to traditional chained equating. Figure 7 displays the differences of raw-to-raw equatings between chained linear and KE CE linear, and chained equipercentile and KE CE optimal. As can be seen, the KE linear conversion is almost identical to the chained linear conversion, with a difference line of zero across the entire score range. The KE optimal conversion was also very close to the chained equipercentile conversion across most of the score range, and the differences below Score Level 1 are most likely due to the equating error.

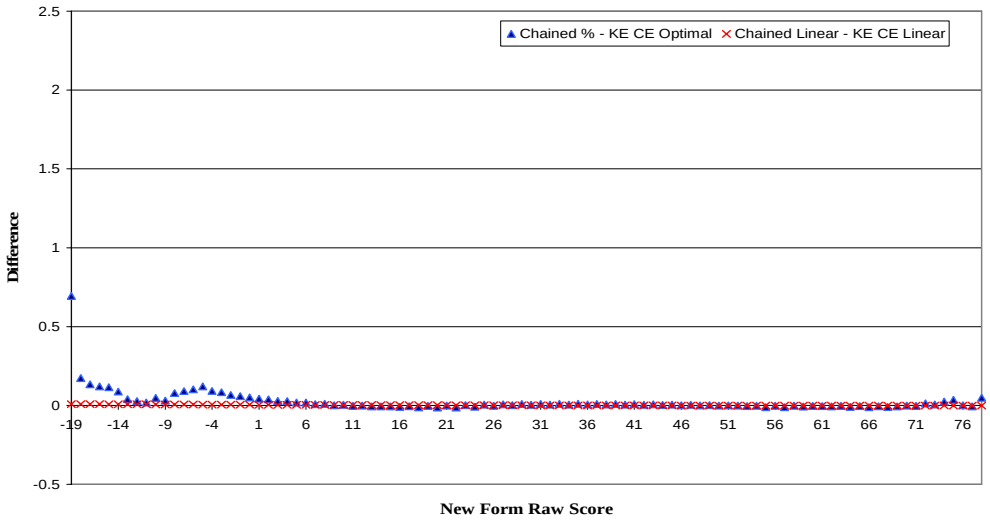

***Figure 7*. Equating to a distant population: raw-to-raw equating between kernel equating and its target approximates in chained equating.**

Table 6 presents the summary statistics of raw-to-raw equating based on KE CE and its target analogue. The means and standard deviations are very close between KE linear and chained linear, as well as between KE optimal and chained equipercentile.

Table 6

Summary Statistics of Raw-to-Raw Conversions of Kernel Equating and Its Analogues in the Equating to a Distant Population

Statistics	Chained equating				Post-stratification equating			
	KE linear	Chained linear	KE optimal	Chained %	KE linear	Tucker	KE optimal	Freq. est.
Mean	35.23	35.23	35.22	35.22	35.78	35.79	35.77	35.78
SD	17.55	17.55	17.61	17.61	17.45	17.48	17.50	17.50
Min.	-18.75	–18.75	-20.17	–19.47	–17.90	–17.97	–19.94	–19.44
Max.	78.62	78.62	77.80	77.85	78.92	78.99	77.74	77.78

Note. KE = kernel equating, freq. est. = frequency estimation.

Comparison of kernel version of post-stratification equating to traditional post-stratification equating. Figure 8 shows the differences of raw-to-raw equatings between KE PSE linear and Tucker and between KE PSE optimal and frequency estimation. While the differences between Tucker and KE PSE linear equatings show a bit of divergence from the zero line, the

differences were still trivial: within ± 0.1. The KE PSE optimal equating also produced results close to those of frequency estimation, with the differences no larger than 0.5.

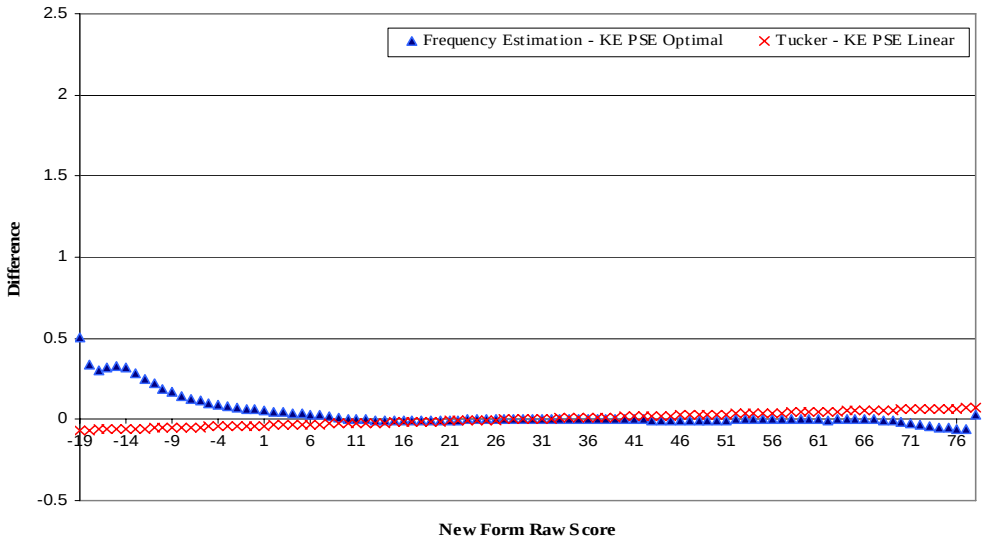

***Figure 8*. Equating to a distant population: differences of raw-to-raw conversions between kernel equating and its target approximates in post-stratification equating.**

Table 6 shows the summary statistics for the post-stratification equating comparisons between KE and its analogue. Under the columns of post-stratification equating, the results of KE PSE linear were very close to the Tucker equating results, and the results of KE PSE optimal were very similar to the frequency estimation results.

For equating to a distant population, each pairwise comparison, KE PSE linear versus Tucker and KE PSE optimal versus frequency estimation, was a good match.

Comparison of KE functions for chained equating and post-stratification equating. Figure 9 displays the SEE for the KE CE linear and the KE PSE linear equatings. Both SEE curves are U-shaped and closely aligned, with the PSE linear SEE values slightly smaller than the CE linear SEE values across the entire score range. The SEED values for CE linear versus PSE linear are plotted in Figure 10. Across the entire scale range, CE measures the new form *X* as being an easier test than PSE does, with the difference line below zero. Further, the difference line lies outside of the ± 2 SEED band except above Score Level 61, suggesting that the differences between the two methods are significantly different from zero and cannot be explained by sampling errors only.

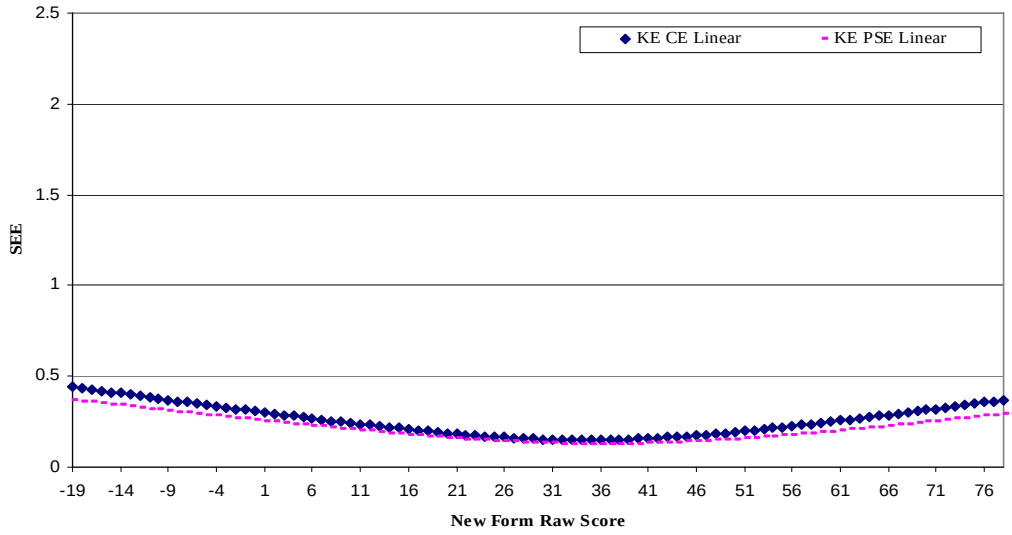

***Figure 9*. SEE for equating to a distant population: KE CE linear versus KE PSE linear.**

Note. SEE = standard error of equating, KE = kernel equating, CE = chained equating PSE = post-stratification equating.

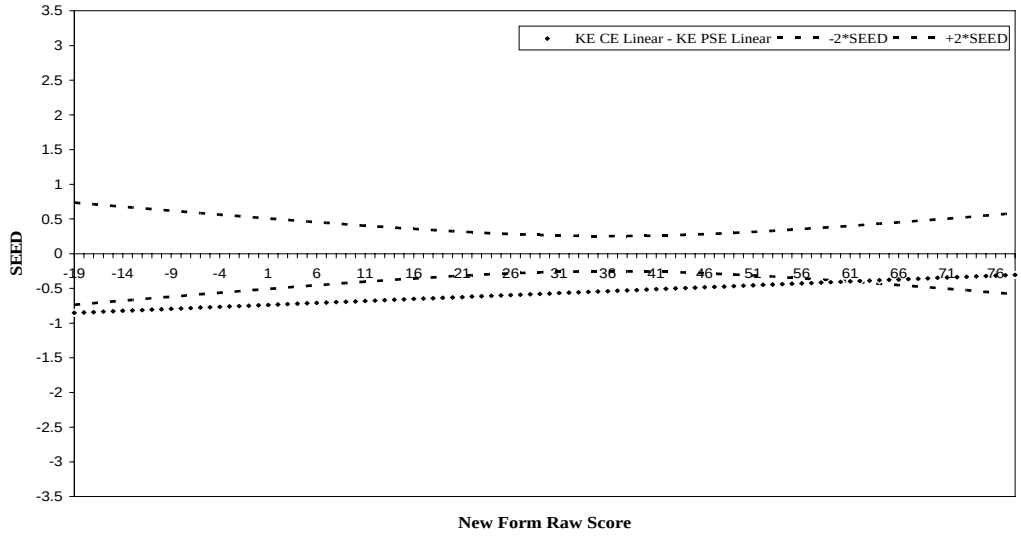

***Figure 10*. SEED for equating to a distant population: KE CE linear versus KE PSE linear.**

Note. SEED = standard error of equating difference, KE = kernel equating, CE = chained equating PSE = post-stratification equating.

Figures 11 and 12 display the differences between KE CE optimal and KE PSE optimal. The SEE values for both, shown in Figure 11, are close to each other. Figure 12 indicates that the SEED is below the zero line until Score Level 71, indicating that CE measures the new form *X* as

being easier than PSE does. The difference line lies outside of the ± 2 SEED band except below Score 6 and above score 66. Above Score 71, CE measures the new form X as being harder than PSE does, but the differences are trivial and lie within the ± 2 SEED band. Overall, when equating to the distant population, CE and PSE no longer produce similar results.

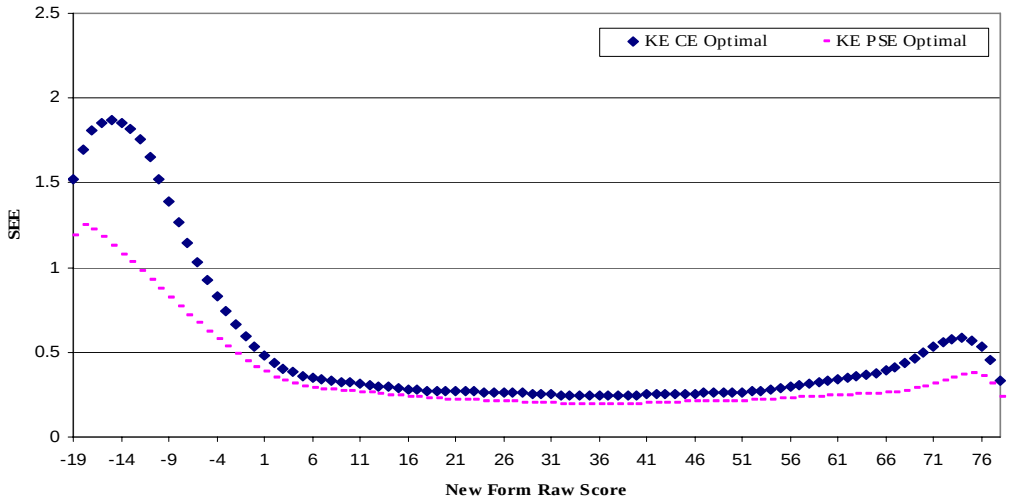

Figure 11. **SEE for equating to a distant population: KE CE optimal versus KE PSE optimal.**

Note. SEE = standard error of equating, KE = kernel equating, CE = chained equating PSE = post-stratification equating.

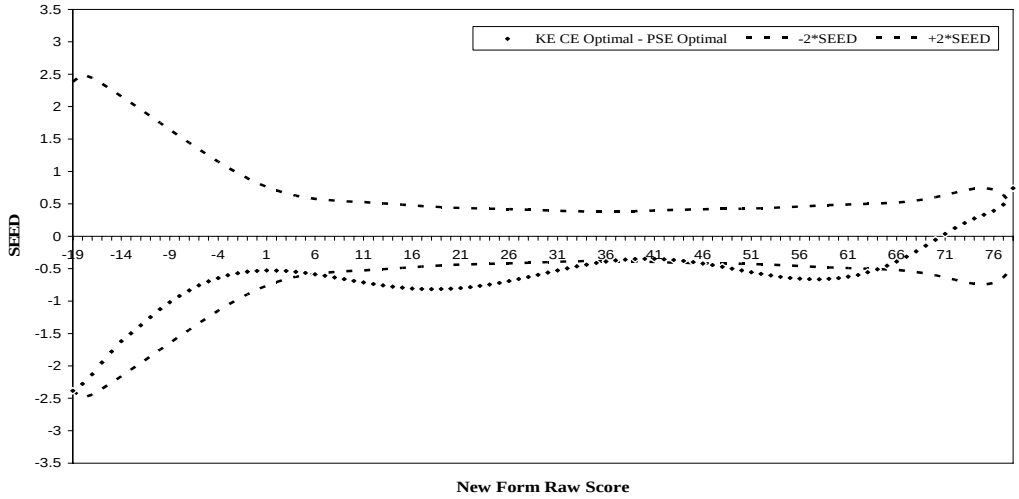

Figure 12. **SEED for equating to a distant population: KE CE optimal versus KE PSE optimal.**

Note. SEED = standard error of equating difference, KE = kernel equating, CE = chained equating PSE = post-stratification equating.

Comparison Between Equating to a Similar Population and to a Distant Population

For each equating method, Table 7 lists the means for equating to a similar population and to a distant population. The pattern is clear when comparing the similar population results to those of the distant population. When equating was conducted to a similar population with the standardized difference of zero, all the equating methods, CE and PSE, KE and classical, gave very similar results. When the new form was equated to a distant population, the separation between CE and PSE methods becomes obvious.

Table 7

Comparison of Means Based on Equating to a Similar Population Versus to a Distant Population

Methods	Similar	Distant
KE CE linear	33.94	35.23
Chained linear	33.94	35.23
KE CE optimal	33.99	35.22
Chained equipercentile	33.99	35.22
KE PSE linear	34.00	35.78
Tucker	33.95	35.79
KE PSE optimal	34.01	35.77
Frequency estimation	34.01	35.78

Note. KE = kernel equating, CE = chained equating, PSE = post-stratification equating.

Discussion

This study applied KE in two scenarios: equating to a very similar population and equating to a quite different population, which is referred to as a distant population in this study. The non-equivalent-groups anchor test (NEAT) design was employed, where the anchor served as an external anchor. The KE results were compared to the results of classical equating methods. Further, the equating results were compared among the methods within the KE family, where the methods were based on different assumptions.

What have we found from this study? First, KE results were comparable to those obtained using classical equating methods. Each pairwise comparison between KE and its

analogue produced very similar results, no matter whether the new form population was similar or different from the old form population. This finding is consistent with the results from other studies (Mao et al., 2005; von Davier et al., 2005).

Second, even though chained equating and post-stratification equating have different assumptions, it is possible that the two types of methods produce the same or similar results when the new form population and the old form population have the same score distribution on the anchor test. von Davier et al. (2004b) proved it theoretically, and the results from equating to the **s**imilar population confirm this theory. They are also consistent with previous research on classical methods: When an anchor test is a miniature of the total test form and is administered to similar groups taking an old and new form, equating methods tend to give similar results (Kolen, 1990).

Third, when groups taking new and the old forms are quite different from each other, equating methods tend to give different results. When equating to the distant population, CE and PSE produced quite different results, and the differences were significantly different from zero. However, since this study used no equating criterion, we cannot determine whether CE or PSE results were closer to the truth.

When evaluating the equating results obtained by using different methods, KE provides some very useful tools, such as PRE and the SEED, that can be taken into account when making final decisions, such as whether the equating differences are real or just due to errors. Based on the results of this study, it is reasonable to use KE operationally, along with other classical equating methods in testing programs.

For future research, in addition to further empirical work with test data that are not so well behaved, such as those from small samples, we may want to consider which equating method is closer to the "truth" when different methods give different results.

References

Dorans, N. J., Liu, J., & Hammond, S. (2004, April). *The role of the anchor test in achieving population invariance across subpopulations and test administrations.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

ETS. (2006). GENASYS [Computer software]. Princeton, NJ: Author.

ETS. (2006). 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 Technical Rep. No. TR-87-79). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Research Rep. No. 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. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, *3*(1), 97–104.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York: Springer-Verlag.

Mao, X., von Davier, A. A. , & Rupp, S. (2005). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS data.* Manuscript in preparation.

von Davier, A. A., Holland, P. W., Livingston, S., Casabianca, J., Grant, M., & Martin, K. (2005). *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, Canada.

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 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.