Research Report

An Evaluation of Kernel Equating: Parallel Equating With Classical Methods in the SAT Subject Tests™ Program

Mary C. Grant

Lilly (Yu-Li) Zhang

Michele Damiano

March 2009

ETS RR-09-06

An Evaluation of Kernel Equating:
Parallel Equating With Classical Methods in the SAT Subject Tests™ Program

Mary C. Grant, Lilly (Yu-Li) Zhang, and Michele Damiano
ETS, Princeton, New Jersey

March 2009

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

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 investigated kernel equating methods by comparing these methods to operational equatings for two tests in the SAT Subject Tests™ program. GENASYS (ETS, 2007) was used for all equating methods and scaled score kernel equating results were compared to Tucker, Levine observed score, chained linear, and chained equipercentile equating results. The results of the kernel chained equatings using a large fixed bandwidth were nearly identical to the results from the chained linear equatings whereas the comparisons of the kernel poststratification equatings using a large bandwidth showed differences from both Tucker and Levine observed score equating results, although for most of the score range those differences were small. Similarly, the differences were small for most of the score range in the comparisons of kernel poststratification equatings and kernel chained equatings using the optimal bandwidth with chained equipercentile equatings.

Key words: Equating, kernel equating, observed score equating

Acknowledgments

The authors would like to thank Alina von Davier for her generous help in coding the kernel equating sections in GENASYS and in interpreting and evaluating the various plots in the kernel equating GENASYS output. Also, thanks go to Tim Moses and Amy Schmidt for reviewing a draft of this report.

Introduction

The typical large scale testing program continually faces the problem of putting new forms of its tests onto the appropriate reporting scale, usually in a fairly short period of time. Therefore, the methods of equating that such a program uses need to be reliably accurate, simple to code, fast to run, and easy to review. A variety of equating methods have been utilized in the SAT Subject Tests™ program to produce scaled scores that are comparable across different test forms and administrations. The usual equating design for the SAT Subject Tests program is the nonequivalent groups with anchor test (NEAT) design. The program currently employs the typical traditional classical linear and nonlinear equating methods: Tucker, Levine observed score, chained linear, and chained equipercentile. The final raw-to-scale conversion is obtained by applying the raw-to-scale conversion used on the reference form to the equated raw scores for the new form. These methods and procedures are all run using GENASYS (ETS, 2007), a comprehensive data analysis system that has standardized coding as well as results and reports that those familiar with them find easy to review. It is a software system that can run the equating methods selected from those available in the system, calculate the raw-to-scale conversions, and store the results on a database.

The kernel method of test equating promises to unify several methods of test equating into a single method while providing new statistical measures. Kernel equating (KE) provides both chained and poststratification observed-score equating methods, which approximate several commonly used traditional equating methods. Since KE can provide both linear and curvilinear equating results, the comparison of curvilinear and linear equatings is possible within the KE methodology. In addition, KE can be applied to all of the usual equating designs. KE also employs the standard error of equating difference (SEED), which is used primarily in a plot that shows the limits of ±2 times the SEED around one equating function and the difference of the other KE function from that criterion line. If the difference line is mostly within the SEED limits, then the function can be considered linear; if it is outside of the line, then the function is curvilinear. Unfortunately, at the time this study was conducted, the SEED was not available in GENASYS (ETS, 2007). Although KE methods have been well developed in theory, thus far they have not been used much operationally.

This study was part of a 2007 initiative to investigate kernel equating in parallel with operational equatings and to evaluate the possible application of KE in testing programs at ETS.

The initiative was directly focused on programs at ETS that currently use observed-score equating methods. As a part of that 2007 research initiative, this study concentrated on the application of the KE method in the operational work of the SAT Subject Tests. This study examined how well the KE method performed compared to traditional equating methods in the testing program.

In this research, KE methodologies were applied in parallel with the classical anchor test equating methods employed operationally for the NEAT design on selected SAT Subject Tests. The equating of a new form for two different tests was examined. Equating results from KE methods were compared with results from traditional linear and nonlinear methods. This study also investigated the conditional standard errors of equating (CSEE) in the different KE methods and compared them to the CSEEs obtained operationally with the traditional equating methods. This study concentrated on the potential application of KE in the operational work of SAT Subject Tests. The main research questions were: How do the KE results compare to the results of the equating methods currently used operationally? And, what are the advantages or disadvantages of the KE method in the SAT Subject Tests operational setting?

Previous Studies

A variety of equating methods have been developed to meet different operational needs and these methods were generated based on different assumptions or theories to produce scores that can be comparable across different test forms. The kernel method of test equating was introduced by Holland and Thayer (1989), who described it as "a new and unified approach to test equating … based on log-linear models for smoothing score distributions and on the kernel method of non-parametric density estimation" (p. i). In 1993, Livingston investigated the accuracy of KE with small samples compared to equipercentile results for both observed distributions and discrete distributions produced by log-linear smoothing. He found that the KEs were much more accurate than the equipercentile equatings using the observed distributions, but they were only slightly more accurate than the equipercentile equatings that used the smoothed distributions.

Thereafter little was published about KE until von Davier, Holland, and Thayer (2004) extensively developed KE. They provided both the theoretical basis for the methodology and examples of KE in commonly encountered testing situations. One of the most distinctive differences between KE and traditional equipercentile equating is that the score distributions are

converted from discrete distributions to continuous distributions in KE by use of Gaussian kernel smoothing as opposed to using a linear interpolation as 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” (p. 45). KE also provides a new measure of statistical accuracy, SEED, which can be used to compare two KE functions, usually the linear and curvilinear results.

More recently, KE has been explored intensively from different perspectives using both operational data sets from various testing programs and simulated data. The research has covered several topics concerning how closely KE approximates classical equating methods under a variety of circumstances. For example, equatings that were conducted operationally using traditional equating methods were repeated using KE (Allspach & Damiano, 2007) and KE results were compared with operational results for testing situations with different test administration designs with different samples sizes and different characteristics (Mao, von Davier, & Rupp, 2006). Other research included using pseudotests with significant differences in difficulty levels for new and reference forms and substantial differences in the ability of new and reference groups (von Davier, Holland, Livingston, Casabianca, Grant, & Martin, 2006), sampling from real data to equate with very small samples (Grant, Zhang, Damiano, & Lonstein, 2006), equating with similar and different populations (Liu & Low, 2007), and equating with varying degrees of presmoothing (Moses & Holland, 2007). There has even been research to expand the scope of KE to include an equipercentile version of Levine (von Davier, Fournier-Zajac, & Holland, 2007).

Kernel equating has been examined and investigated theoretically and practically, and the results from these different perspectives have shown that KE might be a promising method in answering the variety of challenges encountered in operational work. Kernel equating has been shown to be as effective as traditional equating methods. However, even those studies using real data have, so far, been hampered by computer software considerations. Kernel equating has been, for the most part, an experimental methodology, often using stand-alone software. Even as part of a comprehensive and operationally used software, such as GENASYS (ETS, 2007), KE has not been fully developed or tested for use in truly operational circumstances.

Methodology

This study was conceived to conduct KEs completely in parallel with the operational equating of two new test forms in the SAT Subject Tests program during the actual processing time of an administration. The KEs were to be run using the same software system, GENASYS (ETS, 2007), that was used for the operational equatings. In addition, the final evaluation of the effectiveness of KE was to be completed using equated scaled scores rather than equated raw scores.

The two tests chosen for this study were the SAT Subject Test in Mathematics Level 2 and the SAT Subject Test in U.S. History. Both tests were expected to have reasonably large volumes at the time of equating. The mathematics test has 50 items. Previous forms of this test had reliabilities of .90 or higher. The history test has 90 items, and previous forms had reliabilities of .92 or higher. Both tests are five-option multiple choice tests administered with a 1 hour time limit. The test forms were scored with a formula score, a correction for guessing calculated as the number of correct responses minus 0.25 times the number of incorrect responses. Finally these formula scores were rounded to the nearest whole number to create the examinees' raw scores. Possible raw total scores for each form of the mathematics test ranged from -12 to 50, and for the history test, the possible raw total scores ranged from -23 to 90. Examinees taking the SAT Subject Tests select which of the 20 subject tests they will take, and they usually choose those tests that will best highlight their individual strengths. Therefore the ability level of those taking the SAT Subject Tests tends to be higher than that of the SAT Reasoning Test™ population and higher than that of the general college-going population. Consequently, relatively few scores fall below the mean chance score on these tests.

The reference form (the old form in the equating) and the new form shared a set of common items as an internal anchor for the equatings. The total sample for the mathematics test equating used scores from all available records at the time of the operational equating: 13,541 examinees in the new form sample and 14,271 examinees in the reference form sample (see Table 1). The total sample for the history test equating also used scores from all available records at the time of the operational equating: 17,373 examinees in the new form sample and 39,851 examinees in the reference form sample (see Table 2). For the forms used in this study, the mathematics test new form sample was slightly more able and more variable than the reference form sample (see Tables 1 and 3) and the new form was slightly easier than the reference form

with scores that were more variable (see Table 4). The history test new form sample was slightly less able than the reference form sample and slightly more variable (see Tables 2 and 3), while the new form was harder than the reference form with scores that were slightly more variable (see Table 4).

Table 1

Summary Statistics for Anchor and Total Test for the SAT Subject Test in Mathematics Level 2

Form	New form		Reference form	
Sample size	13,541		14,271	
	Anchor	Total	Anchor	Total
Number of items	16	50	16	50
Raw score mean	9.53	27.12	9.20	25.94
Raw score SD	3.98	11.84	3.75	10.86
Observed raw score range	-4 – 16	-6 – 50	-4 – 16	-7 – 50
Skewness	-0.3796	-0.2091	-0.2276	-0.1348
Kurtosis	2.52	2.25	2.44	2.35
Correlation of anchor and total	0.93		0.92	

Table 2

Summary Statistics for Anchor and Total Test for the SAT Subject Test in U.S. History

Form	New form		Reference form	
Sample size	17,373		39,851	
	Anchor	Total	Anchor	Total
Number of items	23	90	23	90
Raw score mean	13.66	51.32	13.99	53.75
Raw score SD	5.38	19.70	5.10	18.05
Observed raw score range	-4 – 23	-8 – 90	-4 – 23	-9 – 90
Skewness	-0.5713	-0.6149	-0.6091	-0.6374
Kurtosis	2.72	2.71	2.88	2.87
Correlation of anchor and total	0.92		0.92	

Table 3

Comparison of Examinee Performance on Anchor Test

Test	SAT Mathematics Level 2	SAT U.S. History
SD in anchor test means $\lvert M_{NA} - M_{RA} \rvert / \sigma_{NARA}$	0.09	0.06
Ratio of anchor test variances $\sigma^2_{NA} / \sigma^2_{RA}$	1.12	1.11

Table 4

Synthetic Group Estimates for Total Scores Using the Tucker Equating Method

Form		SAT Mathematics	SAT U. S. History
New form	Mean	26.65	52.09
	SD	11.56	19.12
Reference form	Mean	26.37	53.43
	SD	11.14	18.31

All of the equatings were conducted using the GENASYS (ETS, 2007) software. Total sample equating with traditional equating methods and KEs were used to obtain both linear and nonlinear raw-to-scale equating results. The traditional equating methods used were Tucker, Levine observed score, chained linear, and chained equipercentile. Kernel equating was run with an input bandwidth set to 0.0 that causes GENASYS to solve for the optimal curvilinear result for both chained equating (CE) and poststratification equating (PSE). It should be noted that the optimal curvilinear result solved has a bandwidth value that is actually greater than 0.0; it is, in fact, impossible to compute a KE function with a bandwidth of 0.0. To obtain linear results in KE, the bandwidth was fixed at approximately 10 times the raw score standard deviation for PSE; this was a bandwidth of 150 for the mathematics test and 200 for the history test. For KE CE, to obtain linear results, a value of 400 was designated as the fixed bandwidth for both the total test and anchor test for both the mathematics test and the history rest. Other parameters in the KE GENASYS coding were set to the default values.

For KE and for the chained equipercentile equatings, the distributions were smoothed using log-linear presmoothing in GENASYS. The moments preserved in the log-linear smoothing for KE were the same as those used for the chained equipercentile method in the

operational equating, which were six univariate moments and one bivariate moment for both the total test and the anchor. No moments were preserved for smoothing the teeth separately even though the two tests used in this study were formula scored. It has been demonstrated that if the teeth were smoothed separately in the presmoothing stage, KE would essentially undo that part of the smoothing in the continuization phase (Moses, 2008). In those situations, KE and traditional methods do show somewhat different results.

The normal operational equatings were performed during the administration processing period. Additional GENASYS (ETS, 2007) requests were coded for the KE of the two test forms selected. These were run within the administration processing window as planned, however, the runs ended abnormally. Problems were encountered in the software execution; consequently, not all of the requested KE results were produced and not all of the requested reports were printed. Therefore, the KEs for this study were postponed until after the administration, by which time the difficulties had been addressed. The KEs for both test forms were completed successfully at that time using the same samples as were used during the on-time classical equatings.

Data Analyses

For both tests, each of the KE results was compared to results from the classical method or methods that were most similar. The results from the KE CE with a large fixed bandwidth were compared with the chained linear equating results. The results from the KE PSE with a large fixed bandwidth were compared with the results from the Tucker equating. They were also compared with the Levine observed score equating results even though these methods are not exactly parallel. At the time of the study, KE in GENASYS (ETS, 2007) had no exact parallel to Levine observed score equating. However, Levine employs a synthetic group, as does Tucker, although the assumptions and procedures for the two methods are different. When the abilities of the two groups are similar with similar standard deviations and when the correlation between anchor scores and total scores are relatively high (see Tables 1, 2, and 3), the results from Levine tend to be close to those of Tucker. Therefore, the results from KE PSE with a large fixed bandwidth were compared to the Levine results as well as to the Tucker results. The results from KE CE with an optimal bandwidth were compared to chained equipercentile results. The obvious comparison for results from KE PSE with optimal bandwidth would be frequency estimation equipercentile equating, however, the SAT Subject Tests program does not use frequency estimation in its current equating procedures. As chained equipercentile is the only curvilinear

method employed by the program, under the circumstances it seemed prudent to compare both KE curvilinear results with results from the one curvilinear traditional method available. Therefore, the results from KE PSE with an optimal bandwidth were also compared to the results from the chained equipercentile equating in spite of the fact that the chained and poststratification methods are based on different assumptions and procedures.

In developing the idea of the *difference that matters* (DTM), Dorans and Feigenbaum (1994) stated that a difference that is less than one-half of a scaled score interval can be considered small enough to ignore. The value of the DTM is the smallest value that would round to a whole scaled score unit. In the case of the SAT Subject Tests, the DTM value is 5 since the scaled scores are reported in intervals of 10. The difference between the unrounded scaled score results at each raw score point from the KE and traditional methods was compared relative to the DTM to determine at what points the KE results were substantially different form those obtained using the traditional equating methods.

The root mean square is a measure of the magnitude of a series of values, such as the difference, at set points, of estimated values and baseline values. The root mean square difference (RMSD) is the square root of the mean of the squares of those differences, represented by

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2},$$

where, in this case, d_i is the difference between the scaled score results from the newly introduced KE method and the scaled score results obtained using the classical equating method for the *i*th new form raw score, and *n* is the number of new form raw scores. The magnitude of the differences between the KE scaled scores and the corresponding classical equating scaled scores were examined.

Because the distributions of scores were negatively skewed, few scores fell below the mean chance score of zero on the two tests. Therefore, the RMSD values were calculated using the scaled score differences for only those raw formula scores at or above the mean chance formula score (i.e., at and above zero). The smaller or the closer to zero the RMSD was, the better a given KE result approximated the corresponding classical equating method result. Since scaled scores are reported in intervals of 10, an RMSD value less than 5 indicated that the KE

method was a good approximation of the classical method, and an RMSD value in the 5 to 10 range indicated that it was a reasonable approximation.

The standard error of equating is defined as random equating error, and it is present when the scores of examinees are sampled from the population. If the whole population were available, then random equating error would not be present. Thus, the amount of random error in estimating the equating relationship becomes negligible as the sample size increases (Kolen & Brennan, 2004). The CSEE is the estimate of the standard error of equating conditioned on the distribution of the raw scores of the new form. The CSEE is used in GENASYS (ETS, 2007) to measure the uncertainty in the estimated equating function of an equated score. The procedures and methods used in GENASYS to compute the CSEE are determined by the different equating methods in the different data collection designs. The descriptions of CSEE procedures and formulas for the different methods can be found in von Davier (2002). The CSEEs produced by GENASYS for the KEs were examined. They were then compared to the CSEEs produced by GENASYS for the same corresponding classical methods that were used in comparing the equating results. It should be noted that the CSEEs used in this study are for the raw-to-scale conversions, not for the raw-to-raw conversions.

Results

Comparison of Equating Lines

The KE equating methods do not map one-to-one with those currently used in equating the SAT Subject Tests. The KE CE with a large fixed bandwidth results were compared with the chained linear results, and the KE PSE with a large fixed bandwidth results were compared with both the Tucker and Levine results. Currently only one classical nonlinear equating method is used in the SAT Subject Tests program: chained equipercentile with presmoothed distributions. Therefore, both the results from the KE CE with an optimal bandwidth and the results from the KE PSE with an optimal bandwidth were compared to the chained equipercentile equating results.

Figures 1 through 6 are all difference plots with the relevant operational result subtracted from the corresponding kernel result. Consequently, the scale on the Y-axis is *difference* in scaled scores.

Figures 1 and 2 clearly show that the KE CE method with large fixed bandwidths for total test and anchor yields results are nearly identical to the traditional chained linear equating method for both tests. For the mathematics test, the largest difference in the scaled score was

0.0005 at the extreme lower end of the scale, while for the history test, the maximum difference was 0.0018, also at the extreme lower end of the scale. Because the KE results, even with a large bandwidth, are not truly linear, there is some, albeit very small, deviation from a straight line.

As expected, for both tests the results from the KE PSE with a large fixed bandwidth more closely resemble the Tucker results than the Levine results (see Figures 3 and 4). For the mathematics test only six scaled scores resulting from the KE PSE were closer to Levine than to Tucker. The differences from Tucker never exceeded the DTM, while the differences from Levine exceeded the DTM for 15 raw score values mostly at the extreme low end of the scale, specifically, -12 to 2, and for two raw scores at the top of the scale, 48 and 49. For the history test, the results from KE PSE with a large fixed bandwidth were clearly more like the results from Tucker than the results from Levine. The differences from the Tucker results did not exceed the DTM at any point, whereas the differences from the Levine results exceeded the DTM at 24 raw score values at the extreme lower end of the scale from -22 to 1. It should be noted that most of these values are below the mean chance score where there were few observed scores.

For the mathematics test the curvilinear results are less clear. Of the values that can be compared, those for raw score values of -6 to 47, 32 scaled score values from the KE PSE with an optimal bandwidth results are closer to the chained equipercentile results than the KE CE with optimal bandwidth results, and only 22 scaled score values from the KE CE with an optimal bandwidth results are closer to the chained equipercentile results than the KE PSE with an optimal bandwidth results. Figure 5 shows that the two KE methods produce results that are similar in the center to top of the raw score scale, from 7 to 45. The KE CE differences from chained equipercentile only exceeded the DTM at five raw score values at the bottom of the raw score scale, from -6 to -2, and one raw score value, 47, at the top of the raw score scale. On the other hand, the KE PSE differences from chained equipercentile exceeded the DTM at 12 raw score values near the bottom of the raw score scale, values of -8, -7, and from -5 to 4.

For the history test, the results from the KE CE with an optimal bandwidth were closer to the chained equipercentile results than were the KE PSE with an optimal bandwidth results except at the lower end of the scale (see Figure 6). The differences of the KE CE with an optimal bandwidth from the chained equipercentile results were greater than those for the KE PSE optimal results primarily for the raw score values of zero and below. But all of the differences for both methods were substantially less than the DTM.

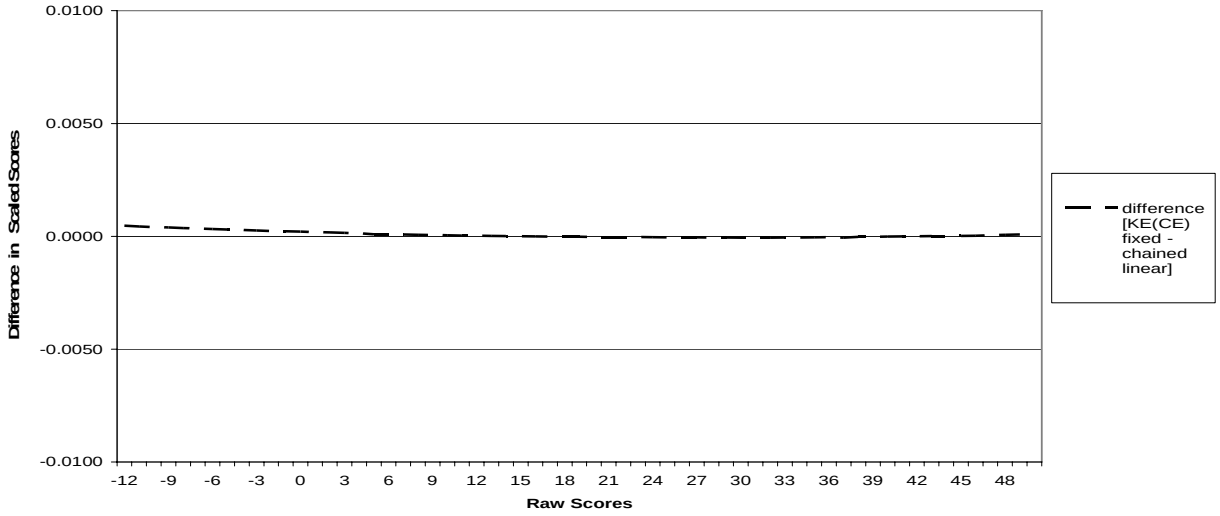

Figure 1. **Difference in scaled scores for equating functions for kernel equating (KE) chained equating (CE) with a large fixed bandwidth and chained linear equating for the SAT Subject Test in Mathematics Level 2.**

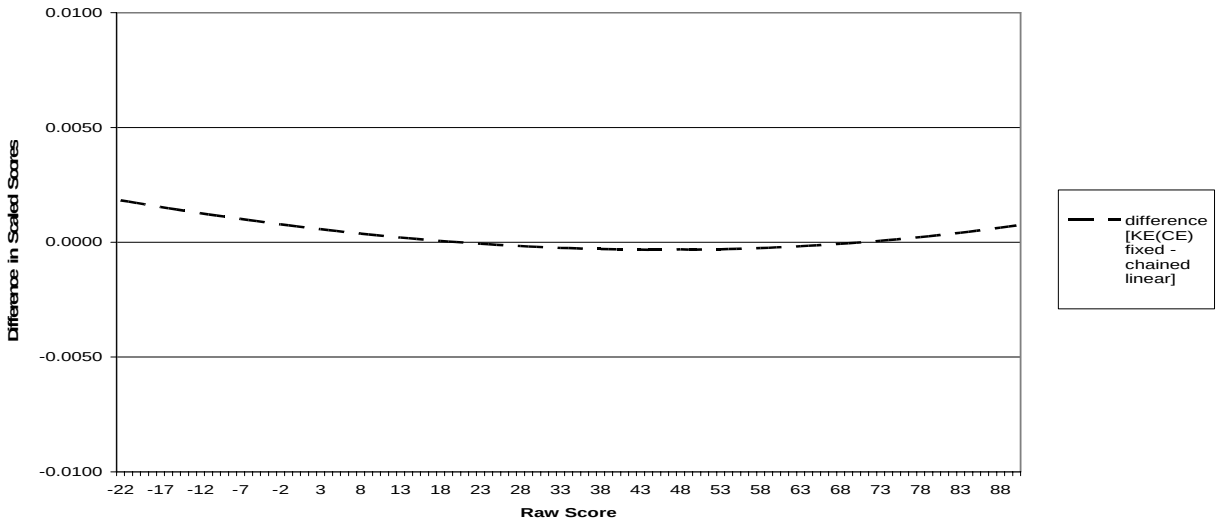

Figure 2. **Difference in scaled scores for equating functions for kernel equating (KE) chained equating (CE) with a large fixed bandwidth and chained linear equating for the SAT Subject Test in U.S. History.**

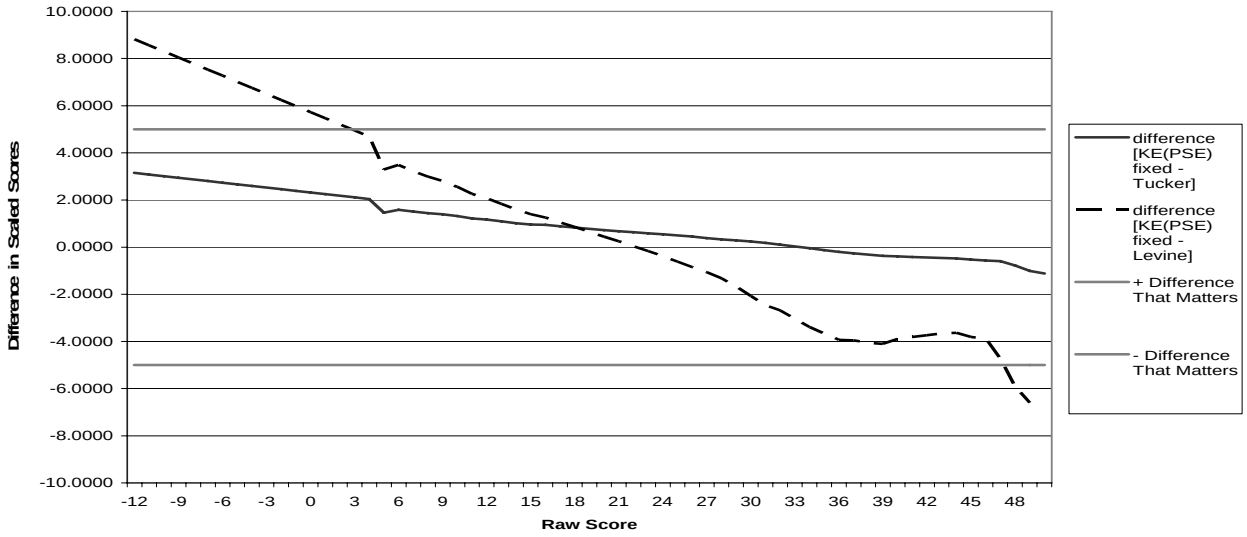

Figure 3. **Difference in scaled scores for equating functions for kernel equating (KE) poststratification equating (PSE) with a large fixed bandwidth and Tucker equating, and difference in scaled scores for equating functions for KE PSE with a large fixed bandwidth and Levine equating for the SAT Subject Test in Mathematics Level 2.**

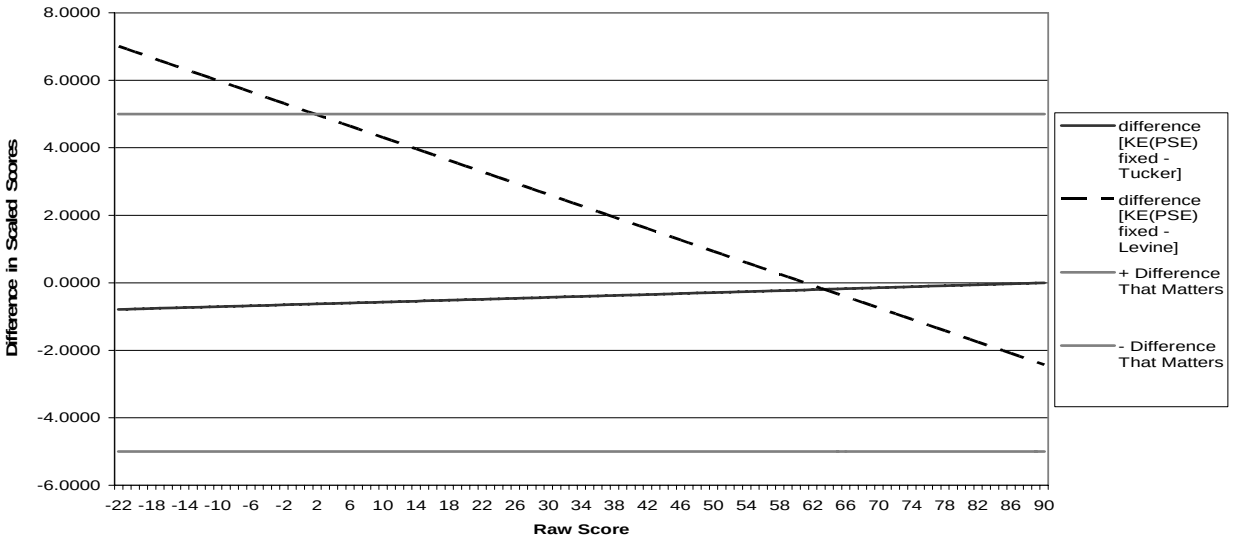

Figure 4. **Difference in scaled scores for equating functions for kernel equating (KE) poststratification equating (PSE) with a large fixed bandwidth and Tucker equating, and difference in scaled scores for equating functions for KE PSE with a large fixed bandwidth and Levine equating for the SAT Subject Test in U.S. History.**

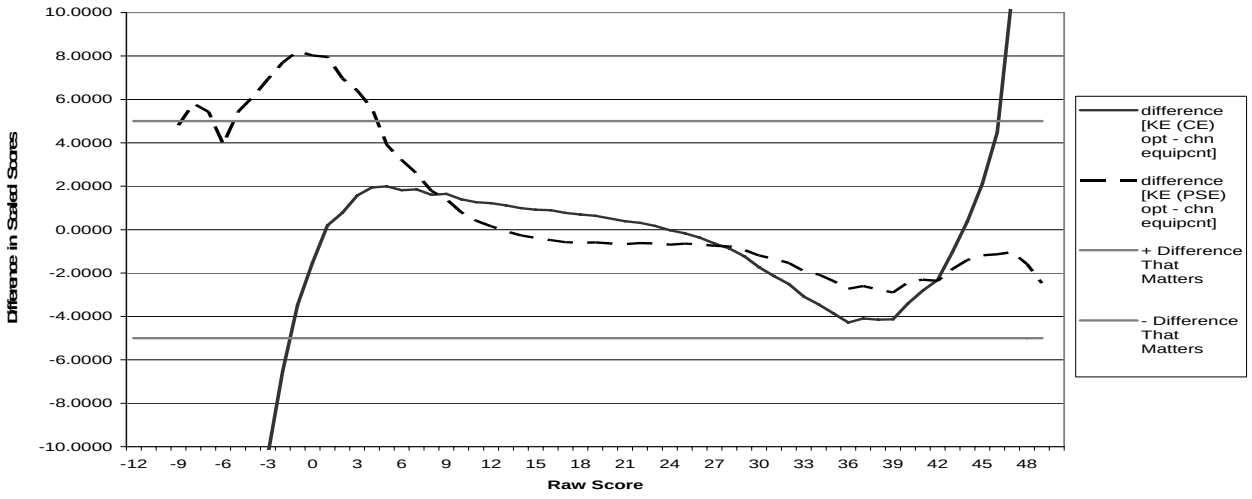

Figure 5. **Difference in scaled scores for equating functions for kernel equating (KE) chained equating (CE) with an optimal bandwidth and chained equipercentile equating, and difference in scaled scores for equating functions for kernel equating (KE) poststratification equating (PSE) ith an optimal bandwidth and chained equipercentile equating for the SAT Subject Test in Mathematics Level 2.**

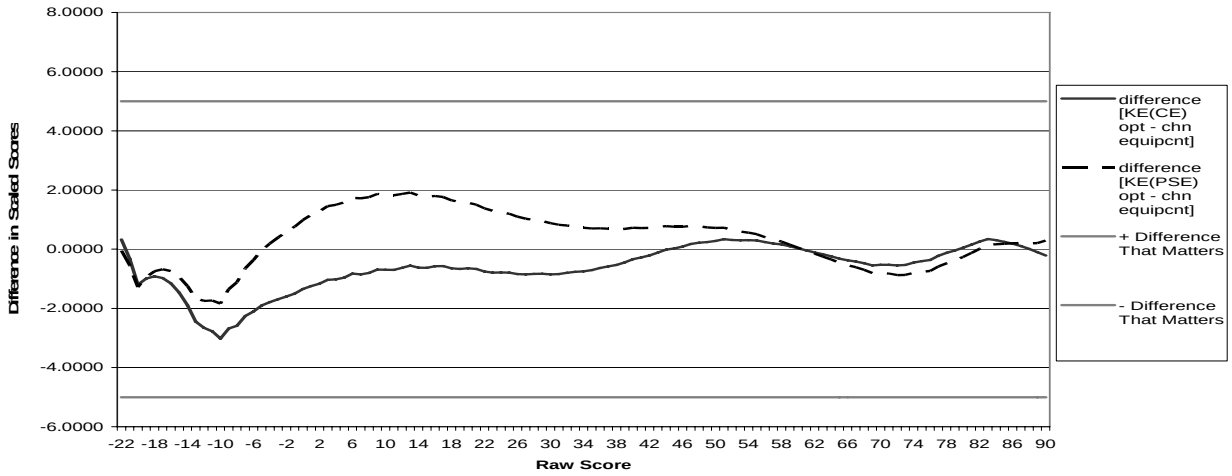

Figure 6. **Difference in scaled scores for equating functions for kernel equating (KE) chained equating (CE) ith an optimal bandwidth and chained equipercentile equating, and difference in scaled scores for equating functions for kernel equating (KE) poststratification equating (PSE) with an optimal bandwidth and chained equipercentile equating for the SAT Subject Test in U.S. History.**

Root Mean Square Difference (RMSD)

Because the distributions of scores were negatively skewed, a few scores fell below the mean chance score of zero in this testing program. For the total samples, only 22 examinees (0.15% of scores) were below zero in the reference form sample and 27 (0.20%) in the new form sample for the mathematics test; 49 examinees (0.12% of scores) were below zero in the reference form sample and 77 examinees (0.44%) in the new form sample for the history test. Since log-linear smoothing methods populate those scores with small fractions of examinees based on the moments preserved in the actual distribution, the equating results for those methods requiring or using the smoothed distributions can be somewhat erratic where little or no real data are available. Therefore, the RMSDs were calculated for raw scores from the mean chance score (zero) and above only, rather than for all of the values.

Table 5 shows the RMSD values for all comparisons for both tests. The RMSD values for the differences between KE CE with a large fixed bandwidth and chained linear showed that the KE was a very good fit with the chained linear results. In fact, those results were close to zero. For the comparisons of KE PSE with a large fixed bandwidth to Tucker and Levine, the RMSDs for both tests indicated that the KE results were a better fit with Tucker than Levine; but the RMSD for both methods for both tests were less than 5.0, and therefore, the KE PSE results were a good fit to both classical methods. For the history test, the results for KE CE with an optimal bandwidth were clearly a better fit with the chained equipercentile results than were the results for KE PSE with an optimal bandwidth where the RMSD for KE CE is 0.57 and the RMSD for KE PSE is 1.01. However, for the mathematics test, the RMSDs showed that the fit of the KE PSE with an optimal bandwidth results and the KE CE with an optimal bandwidth results compared to the chained equipercentile results were essentially the same where the RMSD for KE CE is 2.61 and for KE PSE it is 2.74. In general, the fit for the mathematics test is not as good as the fit for the history test. But, again, all of the RMSD values are less than 5.0. Therefore, for both tests the results for the KE CE and KE PSE with optimal bandwidths were good fits to the chained equipercentile results at and above a raw score of zero.

It should be noted that the Levine and Tucker methods differ from each other more than the relevant continuization method; specifically, the KE PSE with a large fixed bandwidth differs from the traditional PSE method (i.e., Tucker; see Table 5). Therefore, one would expect to see differences between the results from KE PSE with a large fixed bandwidth and Levine. The

comparisons of the results from KE PSE with an optimal bandwidth to the results from the chained equipercentile equatings are somewhat confounded with the differences that are caused simply by the differences in the basic approaches—chained versus poststratification.

Table 5

Root Mean Square Difference (RMSD) for Traditional Method and Kernel Equating (KE) Method for Formula Score Values of Zero and Above Only

Comparison	Test	
	SAT Mathematics Level 2	SAT U.S. History
Chained linear – KE CE fixed bandwidth	0.0001	0.0003
Tucker – KE PSE fixed bandwidth	1.0231	0.3744
Levine – KE PSE fixed bandwidth	3.2801	2.5933
Chained equipercentile – KE CE optimal bandwidth	2.6111	0.5700
Chained equipercentile – KE PSE optimal bandwidth	2.7447	1.0101
Levine — Tucker	2.6050	2.9278
KE CE optimal bandwidth – KE PSE optimal bandwidth	3.0808	1.4499

Note. CE = chained equating; PSE = poststratification equating.

Comparisons of Conditional Standard Errors of Equating (CSEE)

For both tests, the scaled score CSEEs for all of the KE methods are relatively small for new form raw scores where substantial data exist (i.e., for raw scores at and above zero). Also for both tests, the CSEEs for the large fixed bandwidth KEs are smaller than the CSEEs for the optimal bandwidth KEs for both CE and PSE. Figure 7 shows the CSEEs for all four of the KEs for the mathematics test. The CSEEs for the optimal bandwidths, both CE and PSE, are large at the lower extreme of the raw score scale and small throughout the rest of the raw score scale, while the CSEEs for the equatings using a large fixed bandwidth, both CE and PSE, are small for all raw score values. For raw score values at or above a raw score of 6, the differences in CSEE values among the various KE methods are very small. For the history test, the scaled score CSEEs for the two large fixed bandwidth KEs and the two optimal bandwidth KEs have much the same pattern as those for the mathematics test (see Figure 8).

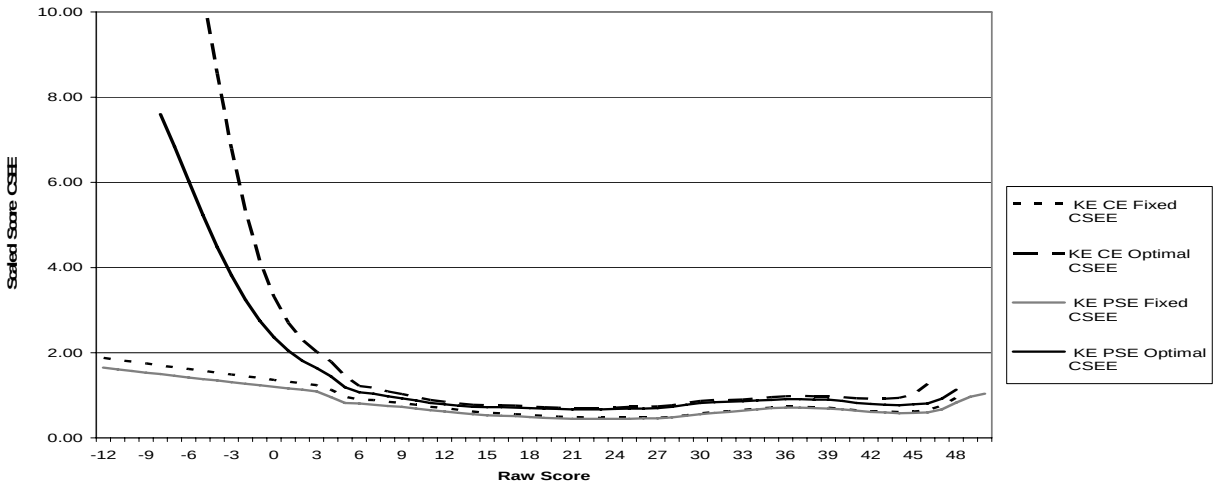

Figure 7. **Conditional standard errors of equating (CSEE) for all kernel equating (KE) methods for the SAT Subject Test in Mathematics Level 2.**

Note. CE = chained equating; PSE = poststratification equating.

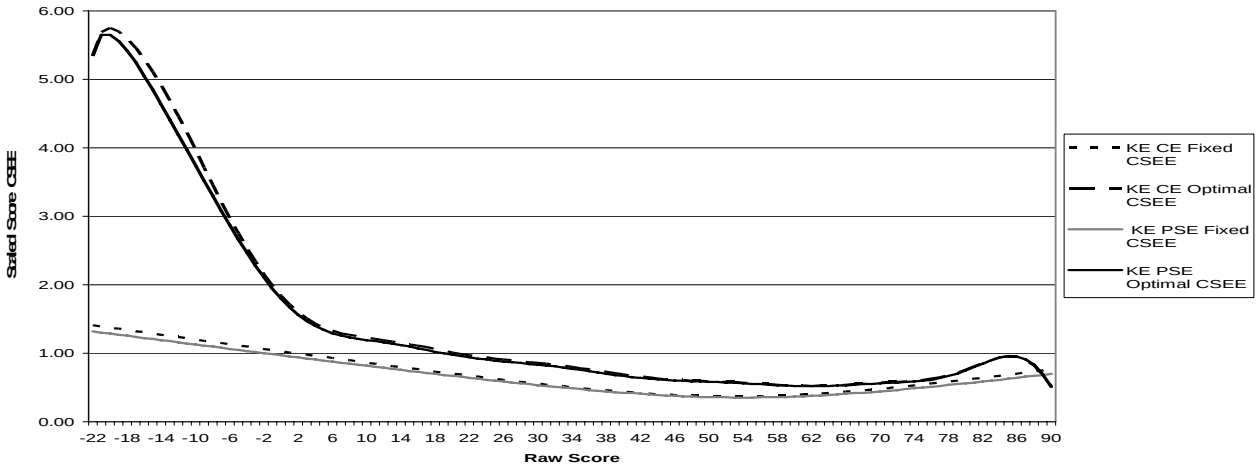

Figure 8. **Conditional standard errors of equating (CSEE) for all kernel equating (KE) methods for the SAT Subject Test in U.S. History.**

Note. CE = chained equating; PSE = poststratification equating.

For both the mathematics test and the history test, the CSEEs for the chained linear and KE CE with a large fixed bandwidth are nearly identical, as were the equated scores. For the mathematics test (see Figure 9), the plots of the CSEEs are a somewhat wavy U-shape reaching

1.9 at the lowest raw score; but for both equating methods, they are below 1.0 for most of the raw score scale. The plots of the CSEEs for the history test (see Figure 10) are somewhat U-shaped with the values for CSEE reaching 1.4 at the lowest raw score. The CSEEs are below 1.0 for raw scores of zero and above for both equating methods.

For the mathematics test, the CSEEs for the poststratification linear equatings (i.e., Tucker and KE PSE with a large fixed bandwidth) and for the Levine equating, all have wavy U-shaped plots (see Figure 11). The CSEEs for the mathematics test are higher at the lower end of the raw score scale than at the top the raw score scale. The CSEEs for Tucker and the KE PSE with a large fixed bandwidth equatings are nearly the same for the whole scale. The CSEEs for the Levine equating are slightly higher by less than 0.5 throughout the scale. All of CSEE values for these equatings are below 2.0, except for the lowest three values for Levine.

Figure 12 shows the CSEEs for the poststratification linear equatings (i.e., Tucker and KE PSE with a large fixed bandwidth) and for the Levine equating for the history test. All of these methods have the typical U-shaped plot with a more pronounced upward tail at the lower end of the raw score scale than at the upper end of the raw score scale. The CSEEs for Tucker and the KE PSE with a large fixed bandwidth equating are nearly the same for most of the raw score scale. However, some separation occurs in the top third of the raw score values. The KE PSE with a large fixed bandwidth has lower CSEE values in this area. The CSEEs for Levine are slightly higher (i.e., less than 0.2 higher) throughout the raw score scale. All of CSEE values for the three equating methods are below 2.0. For the mathematics test, the chained equipercentile equating has very large values for CSEEs at the bottom of the scale, but they fall below 2.0 at a raw score of 9. The KE CE with an optimal bandwidth CSEE values are greater than 10.0 at the bottom, however, the values for the CSEEs are below 1.0 for most of the raw score scale. The KE PSE with an optimal bandwidth CSEEs are similar to those for KE CE with an optimal bandwidth, with a value greater than 7.0 at lower end of the raw score scale and values of less than 1.0 for most of the raw score scale (see Figure 13).

For the history test, the CSEEs for the chained equipercentile, KE CE with an optimal bandwidth, and KE PSE with an optimal bandwidth equatings are much like those seen for the mathematics test. While the CSEE values are high at the lower end of the raw score scale, the CSEE values for the chained equipercentile equating fall below 2.0 for most of the raw score

scale, and the CSEE values for the KE CE with an optimal bandwidth and KE PSE with an optimal bandwidth equatings fall below 1.0 for most of the raw score scale (see Figure 14).

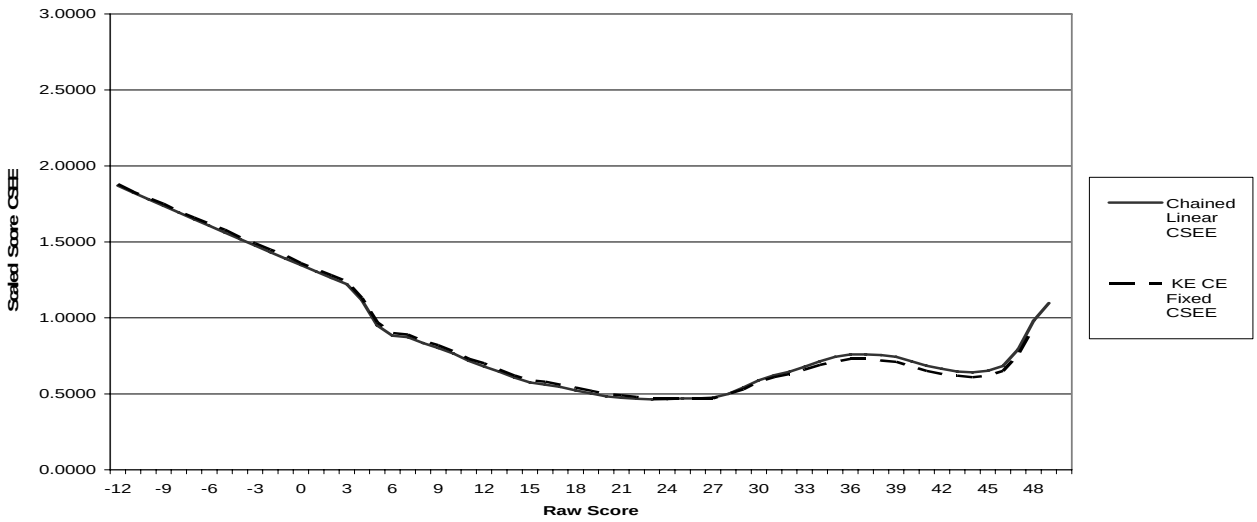

Figure 9. **Comparison of conditional standard errors of equating (CSEE) for chained linear equating and kernel equating (KE) chained equating (CE) with a large fixed bandwidth for the SAT Subject Test in Mathematics Level 2.**

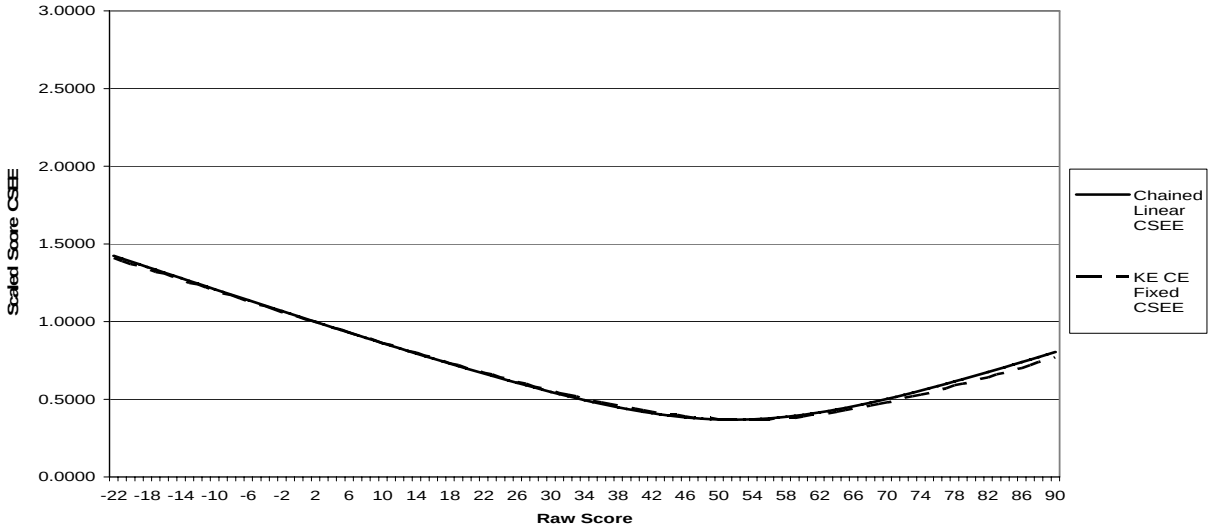

Figure 10. **Comparison of conditional standard errors of equating (CSEE) for chained linear equating and kernel equating (KE) chained equating (CE) with a large fixed bandwidth for the SAT Subject Test in U.S. History.**

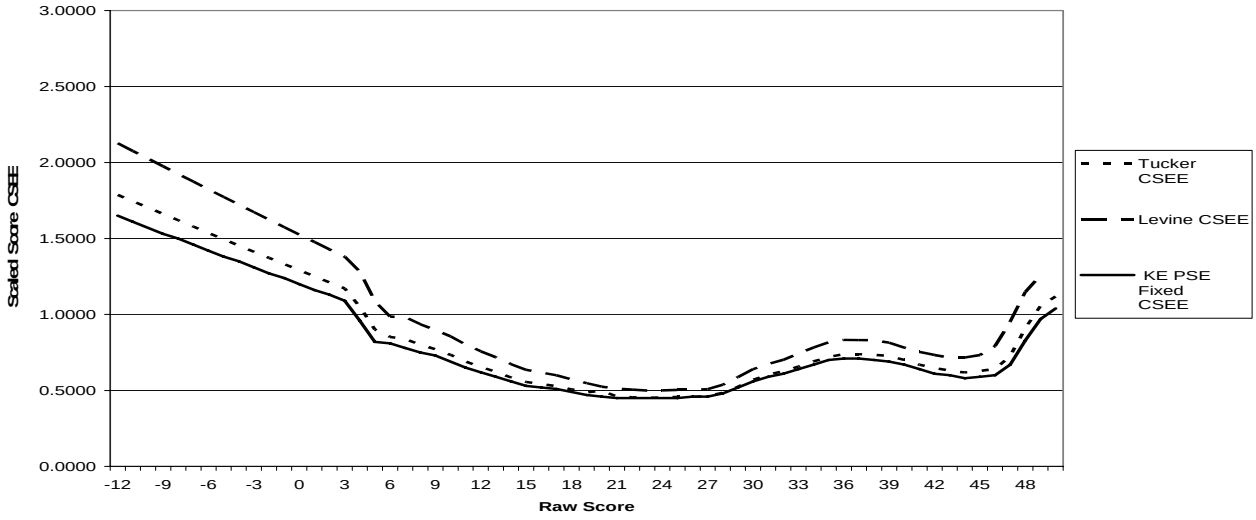

Figure 11. **Comparison of conditional standard errors of equating (CSEEs) for Tucker equating, Levine equating, and kernel equating (KE) poststratification equating (PSE) with a large fixed bandwidth for the SAT Subject Test in Mathematics Level 2.**

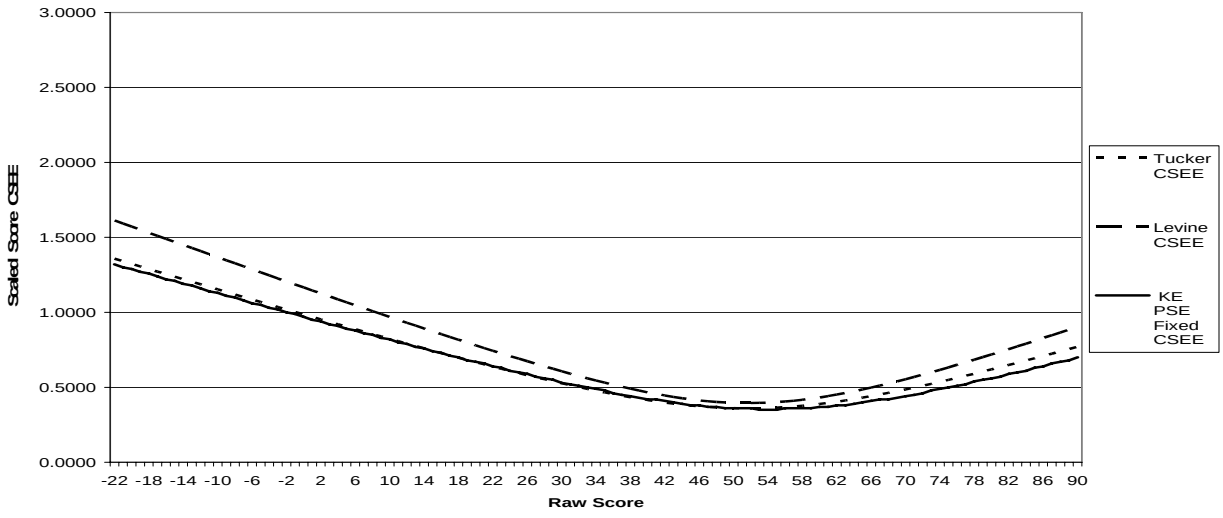

Figure 12. **Comparison of conditional standard errors of equating (CSEEs) for Tucker equating, Levine equating, and kernel (KE) poststratification (PSE) with a large fixed bandwidth for the SAT Subject Test in U.S. History.**

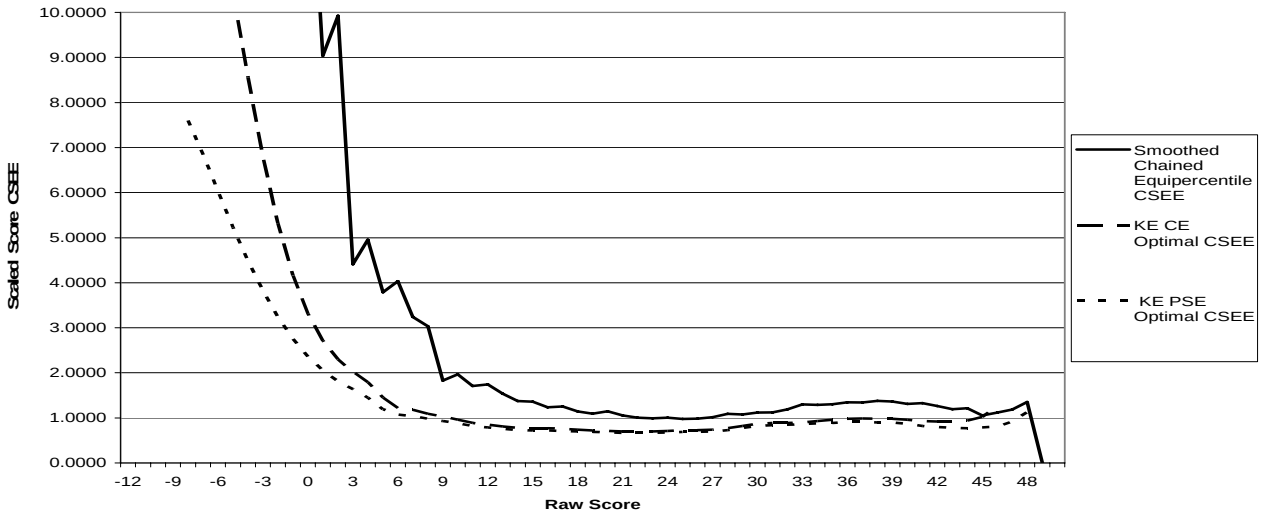

Figure 13. **Comparison of conditional standard errors of equating (CSEEs) for chained equipercentile equating, kernel (KE) chained equating (CE) with optimal bandwidth, and KE PSE (poststratification equating) with optimal bandwidth for the SAT Subject Test in Mathematics Level 2.**

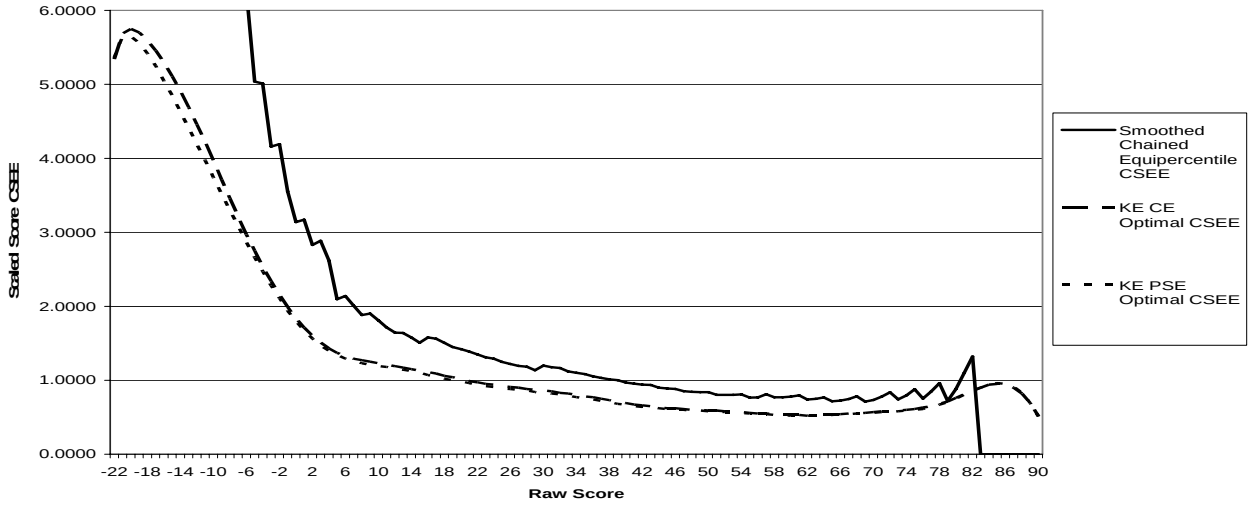

Figure 14. **Comparison of conditional standard errors of equating (CSEEs) for chained equipercentile equating, kernel equating (KE) chained equating (CE) with optimal bandwidth, and KE poststratification (PSE) with optimal bandwidth for the SAT Subject Test in U.S. History.**

Summary and Conclusions

Clearly the KE method that is most nearly parallel with a classical method is the kernel chained equating with a large fixed bandwidth compared to the chained linear equating. Except for running a few more examples to confirm this close match, little needs to be done to aid in interpreting KE CE with a large fixed bandwidth.

In this study, the KE PSE with a large fixed bandwidth conversion matched the Tucker result more closely than the Levine result, as was found in previous studies (e.g., see von Davier et al., 2006). This result was expected since the KE PSE with a large fixed bandwidth equatings and Tucker are based on similar assumptions. The difference between KE PSE with a large fixed bandwidth from Levine was relatively small because the abilities of the two groups were similar with similar standard deviations and the correlation between anchor scores and total scores were relatively high. Therefore, the results from Tucker and Levine were not very different from one another. In addition, the difference of the difficulties for the new and reference forms were not very large. Since the results from the Tucker and Levine equating methods tend to be different when group differences are large, the closeness of the KE PSE with a large fixed bandwidth result to Levine is not expected to hold in situations where large group differences occur. On the other hand, the closeness of the KE PSE with a large bandwidth result to the Tucker result may pose a problem if the KE PSE with a large fixed bandwidth demonstrates the same bias that Tucker shows when the abilities of the groups are substantially different from one another. Further investigation is needed on how KE PSE with a large fixed bandwidth behaves relative to Tucker and Levine when the reference and new form groups are different so that psychometricians can learn how to interpret and evaluate the KE results.

The SAT Subject Tests program has only used one nonlinear equating method operationally; that method is chained equipercentile. With KE, now two methods, using an optimal bandwidth for both CE and PSE, are equally easy to perform in GENASYS (ETS, 2007). Both methods yield results that are not very different from the chained equipercentile method and, at least in this study, not very different from each other (see Table 5). However, only experience with many equatings will educate the psychometrician in evaluating and distinguishing between those two nonlinear equating results.

The KE CE method with an optimal bandwidth and the KE PSE method using a large fixed bandwidth produced results similar to the respective traditional equating methods. The

kernel chained method using a large fixed bandwidth produced results nearly identical to the respective traditional equating method. For both of the tests used in this study, the forms were of similar difficulty, and the groups used in the equatings were of similar ability. Further trials need to be conducted to determine whether the relationships hold when test forms and/or groups are different.

The CSEEs for the KEs were similar to the CSEEs for the respective traditional equatings, with the CSEEs for the KE CE with a large fixed bandwidth being nearly identical to those of the chained linear equatings. The CSEEs for the KE equatings were almost always less than those for the traditional equatings. However, the difference between them was not large except at the bottom of the raw score scale where no data existed. A noticeable advantage for the KEs with optimal bandwidth was that the CSEEs did not go to ridiculously high values at the bottom of the raw score scale.

An important aspect of transitioning to KE is a thorough understanding of how to interpret and compare the final raw-to-scale conversion lines by the psychometricians. The psychometricians need to develop a thorough understanding of how to interpret the KE results for all types of testing situations, such as small sample tests with very skewed distributions. Staff at all levels need to become familiar with the coding options and output so that when introduced operationally to any testing program the KE methods do not delay the processing and therefore delay the reporting of scores. In classical equating, this testing program has used Tucker, Levine, chained linear, and chained equipercentile equating methods. The requirements for the use of each of these methods were thoroughly known so that the equating decisions followed from what was observed in the data and equating output. At the time of this study, in KE in GENASYS (ETS, 2007), two distinct choices were available for both linear and nonlinear equatings. More familiarity with KE results, some of which comes from experience, will be necessary when the psychometrician must work with KE in an operational setting.

KE provides a measure of the accuracy to the estimated difference between two equating functions, referred to as SEED. It is used primarily in a plot that shows the limits of ±2 times the SEED around one equating function and the difference of the other KE function from that criterion line. If the difference line is mostly within the SEED limits, then the function can be considered linear; if it is outside of the line, then the function is curvilinear. This additional evaluation tool used as an integral part of KE is a primary reason for transitioning to KE operationally. The most

disappointing aspect of running KE in GENASYS (ETS, 2007) was the absence of the SEED, which has not yet been written into GENASYS. How having the SEED available for interpretation would affect the experience of working with KE cannot be attested here.

A final question that should be addressed before KE is implemented operationally is: What does KE offer to the equating process? Instituting KE as an additional procedure will increase the information obtained from the equating little, if any, as long as the close relationships between the KE equatings and the respective traditional methods hold under all circumstances. Simply adding KE to the existing operational methods would be more of a redundancy than an addition of information and choice. Therefore, it is as a replacement for traditional equating methods that KE seems to have merit. Perhaps the addition of the SEED to improve the ability of the psychometrician to evaluate the equating functions produced would make KE more appealing. Otherwise, if the results are essentially the same as those that can be obtained using the traditional classical methods, then why go to the effort and expense of switching to KE?

References

Allspach, J. R., & Damiano, M. (2007, April). *Evaluating kernel equating and traditional equating methods with real test data: How comparable is kernel equating?* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT*® (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

ETS. (2007). GENASYS [Computer software]. Princeton, NJ: Author.

Grant, M. C., Zhang, Y-L., Damiano, M., & Lonstein, L. (2006, April). *An evaluation of the kernel equating method: Small sample equating in non-equivalent groups.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (ETS Research Rep. No. RR-89-06). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Research Rep. No. RR-89-07). Princeton, NJ: ETS.

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

Liu, J., & Low, A. C. (2007). *An exploration of kernel equating using SAT® data: Equating to a similar population and to a distant population* (ETS Research Rep. No. RR-07-17). Princeton, NJ: ETS.

Livingston, S. A. (1993). *An empirical tryout of kernel equating* (ETS Research Rep. No. RR-93-33). Princeton, NJ: ETS.

Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS™ data* (ETS Research Rep. No. RR-06-30). Princeton, NJ: ETS.

Moses, T. (2008, June). *Smoothing and equating.* Paper presented at the Equating Special Interest Group seminar series, ETS, Princeton, NJ. [

Moses, T., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (ETS Research Rep. No. RR-07-15). Princeton, NJ: ETS.

von Davier, A. A. (2002). *GENASYS addendum to statistical procedures and methods guide, Version 1*. Princeton, NJ: ETS.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating* (ETS Research Rep. No. RR-07-14). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method. A special study with pseudo-tests constructed from real test data* (ETS Research Rep. No. RR-06-02). Princeton, NJ: ETS.

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