Research Report

The Impact of Anchor Test Length on Equating Results in a Nonequivalent Groups Design

Kathryn L. Ricker

Alina A. von Davier

Research & Development

December 2007
RR-07-44

The Impact of Anchor Test Length on Equating Results in a Nonequivalent Groups Design

Kathryn L. Ricker and Alina A. von Davier
ETS, Princeton, NJ

December 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 explored the effects of external anchor test length on final equating results of several equating methods, including equipercentile (frequency estimation), chained equipercentile, kernel equating (KE) poststratification PSE with optimal bandwidths, and KE PSE linear (large bandwidths) when using the nonequivalent groups anchor test (NEAT) design. This study used pseudotests constructed of item responses from a real operational test. The equating methods were evaluated using an equating criterion. Conditional differences between the criterion scores and equated scores, and root mean square error of the differences (RMSE) were used as measures to compare the methods to the criterion equating, which in this study is an equivalent groups (EG) equipercentile equating function. The results indicate that bias tended to increase in the conversions as the anchor test length decreased, but the KE PSE with optimal bandwidths and equipercentile (frequency estimation) methods were less sensitive to this change than the other methods. The KE PSE linear method with large bandwidths performed poorly compared to the criterion across all anchor test lengths.

Key words: Kernel equating, NEAT design, equipercentile equating, equating bias, difference that matters, anchor test length

Acknowledgments

This paper focuses on some results from a larger study that was conducted by von Davier, Holland, Livingston, Casabianca, Grant, and Martin (2005). It was originally presented at the annual meetings of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), held between April 10 and April 12, 2006, in San Francisco, CA.

Introduction

In practical equating situations, the most common equating design is the nonequivalent groups anchor test (NEAT) equating design, which uses a set of common anchor items to adjust for differences in test difficulty. Use of this design is critical to equating test forms in most large-scale testing programs, because test forms are not equivalent (i.e., parallel), nor can it always be assumed that population parameters will remain stable over time; thus the samples of test takers cannot be assumed to be equivalent over time.

The anchor test ideally acts as a surrogate, both substantively and statistically, for either the remaining items in the test forms (internal anchor) or the items on which the test form scores are based (external anchor; Cook & Peterson, 1987). In general, a longer anchor test is considered desirable, because it is more reliable and it tends to generate fewer random equating errors (Budescu, 1985).

Practitioners can choose from many equating methods for the NEAT design. These include popular equating methods, such as the equipercentile frequency estimation and chained equipercentile methods (among many others), as well as newer methods such as kernel equating (KE). Kernel equating is a variation on classical equipercentile observed-score equating that employs a Gaussian kernel to continuize the discrete observed score distributions (von Davier, Holland, & Thayer, 2004).

The purpose of this study is to explore the effect of external anchor test length on final results of several equating methods, including equipercentile (frequency estimation), chained equipercentile, KE poststratification (PSE) with optimal bandwidths (which emulates the frequency estimation equating), and KE PSE linear (KE PSE with large bandwidths, which emulates the 1982 Braun and Holland linear equating method). The KE version of chained equating was not included in this study due to software limitations. Especially of interest was the behavior of the KE functions when the length of the external anchor was varied, because no empirical reports currently exist regarding the relationship between KE equating performance and anchor test length.

Real data taken from operational testing results were used in this study. Items from actual operational test forms were selected to create two pseudotest forms and anchor sets of varying lengths (see the appendix). Creating the pseudotest forms provided an opportunity to use real data in a systematically controlled way.

In the NEAT design, the two most important test scores from Forms *X* and *Y* (the forms to be equated) are each observed only on population *P* (the group who takes Form *X*) or only on population *Q* (the group who takes Form *Y*), but not on both. However, an anchor test *A* is taken by the groups of examinees from both populations. Thus, *X* and *Y* are not both observed on the target population *T*, and *A* is observed in both, and therefore, *A* will be used to adjust for the differences in overall difficulty between *X* and *Y* (see Table 1). Assumptions must be made in order to overcome the lack of complete information in the NEAT design. Any equipercentile equating method used with the NEAT design makes acceptable and sufficiently strong assumptions that allow one to find values for the cumulative distribution functions (cdfs) of *X* and *Y* in population *T*, $F_T(x)$ and $G_T(y)$, respectively. Similarly, any linear equating method for the NEAT design relies on untestable assumptions about the missing data in order to estimate the means, variances, and eventually the covariances of the variables *X*, *Y*, and *A*.

Table 1

Research Design

Target population	Original populations	Nonequivalent anchor test (NEAT)			Equivalent groups (combined group)	
		X	A_i	*Y*	*X*	*Y*
T[a]	P	√	√	(√)	√	√
	Q	(√)	√	√	√	√

Note. Shaded boxes indicate existing data that were not used for equating. Because both test forms were created from one original test form, both groups had data for both test forms.
[a]Under the equivalent groups design, the target population *T* was calculated by combining populations *P* and *Q* using the formula $T = wP + (1 - w)Q$, where w is the proportional weight for population *P*.

In other equating and test linking designs, such as equivalent groups or single group designs, the target population is simply the group from which the examinees are sampled. In those cases, we may estimate $F_T(x)$ and $G_T(y)$ directly from the observed data. In the NEAT design, however, assumptions that are not directly testable must be added to the mix.

Kernel Equating Framework

KE is an equipercentile equating procedure in which the score distributions to be equated are converted from discrete distributions to continuous distributions by using a normal (Gaussian) kernel as opposed to using linear interpolation, as in the traditional equipercentile equating method (Holland & Thayer, 2000; von Davier et al., 2004). The KE framework consists of five procedural steps: presmoothing the data using loglinear models; computing the marginal score probabilities for *X*, *Y*, and eventually for *A* (for chained equipercentile); continuizing the frequency distributions using the Gaussian kernel; computing the equipercentile equating function using these continuous distribution functions; and eventually, computing the accuracy measures such as the standard errors of equating (SEE) and the standard errors of equating differences (SEED), as shown in von Davier et al. (2004).

The main difference between the KE method and the traditional equipercentile method depends on the continuization step. Kernel equating was devised originally as a solution to a problem arising from the equipercentile definition of equated scores. By this definition, Score *x* on Form *X* and Score *y* on Form *Y* are equated in a population of test-takers if and only if they have the same percentile rank in that population. But in the real world of educational testing, it is rare to find a score on Form *Y* that has exactly the same percentile rank in the test-taker population as Score *x* on Form *X*. This problem arises because the score distribution on a given test form is discrete. The KE method replaces the discrete score distributions with continuous distributions and then equates scores on the continuous distributions.

Basically, by adding a continuous random variable ***V*** distributed *N* (0, 1), the discrete random variables ***X*** and ***Y*** are transformed into continuous variables $X(h_X)$ and $Y(h_Y)$ as:

$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_X$ and $Y(h_Y) = a_Y(Y + h_Y V) + (1 - a_Y)\mu_Y$, respectively.

In the above formulas, h_X and h_Y can be any positive number. They are the bandwidths of the continuous distributions for each discrete score; μ_X and σ_X^2 denote the mean and variance of variable ***X*** over target population *T*; $a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2}$ is an adjusting constant that insures that the mean and variance of $X(h_X)$ are the same as those of *X*. Since the variable ***V*** has a continuous normal distribution, it is obvious that $X + h_X V$ will be continuous, and so is $X(h_X)$. Similar notations are used for $Y(h_Y)$.

The selection of h_X ($or\ h_Y$) determines the equating method. The KE PSE optimal (equipercentile) equating method selects optimal values for h_X (or h_Y) automatically by minimizing the difference between the probability distributions of *X* (or *Y*) before and after continuization (and by using some additional penalty functions—see von Davier et al., 2004). The KE PSE equating method approximates a linear method by using large bandwidths, values that are usually larger than 10 times the standard deviation of the distribution to be continuized.

Equating Criterion

The evaluation of any equating method requires the use of one (or several) equating design(s) where the equating criterion is known (Harris & Crouse, 1993). In practice it is very difficult to find a known criterion for equating, particularly in a study where real data are used. Here we use extended data from pseudotests that were initially constructed and described in von Davier et al. (2005). A single long test form was used at two different administrations to two nonequivalent groups; the items of this test were used to construct two different shorter forms that differed in difficulty and three anchor tests that differed in length. Using this design, we have data from two forms that can be equated using a NEAT equating design and can also be equated using an equivalent group design (EG) in the combined group. The study design is summarized in Table 1.

To provide a criterion for the accuracy of the anchor equating methods, we used the classical equipercentile equating method to equate the presmoothed (with log-linear models) distributions of scores on the forms in the EG design (i.e., the combined group of examinees from the two test administrations) as the basis for our evaluation of the equating results from the other conditions.

It is recommended to explicitly define the target population *T* for a given equating design. In this study where we are interested in evaluating the equating methods in a NEAT design, *T* is assumed to be a mixture of *P* and *Q,* in which *P* and *Q* are regarded as nonoverlapping, nonequivalent subpopulations, which make up *T*. *P* and *Q* are given weights that sum to 1, which could be proportional to their relative population sizes. This is denoted by $T = wP + (1 - w)Q$, where *w* is the relative weight of population *P* in population *T*. The criterion equating should be done on the same population as was used for the equating methods we are interested in evaluating. Therefore, the criterion equating design, the EG design, was computed by pooling the data from the two administrations, insuring that the target population *T* is of the form $T = wP +$

$(1 - w)Q$, with the w determined by the relative size of the samples from P and Q (i.e., $w = n_P/(n_P + n_Q)$, where n_P and n_Q are the sample sizes of the samples from P and Q, respectively). The score distributions computed for P and Q separately are weighted by w and $(1\text{-}w)$ to obtain distributions of these same quantities for T. See Table 1 for the illustration of the NEAT design and of the EG design obtained from combining the groups.

von Davier et al. (2005) investigated whether, in order to define an equating criterion, one should check if the criterion equating in the combined group is the same as the equatings inside each of the groups. The results of these additional equatings are given in Appendix B in von Davier et al. (2005) and show that the equatings are similar in the score range where the data are available. However, the authors consider that "these analyses check a population invariance assumption and cannot influence the choice of the criterion. The choice of the equating criterion is based on a decision about the appropriate target population and eventually, about the appropriate shape of the equating function." (von Davier et al., 2005, p.10).

Assessing Equating Methods Relative to the Criterion

The effects of external anchor test length will be examined in this study through measures of conditional differences between the criterion scores and equated scores at each raw scale score point, as well as global measures such as the root mean square error of the differences (RMSE). It is helpful to have guidelines to aid in interpretation of the results of these analyses. One practical guideline is the use of the difference that matters (DTM; Dorans & Feigenbaum, 1994), which has been used in previous equating research (e.g., Ricker & Gierl, 2005; von Davier & Han, 2004). Briefly stated, Dorans and Feigenbaum (1994) defined a DTM as any score difference that would make a difference in score reporting once scores were rounded. In this study, where only raw scores are being considered, a DTM is defined as any score difference that is equal to or greater than 0.5.

Method

Instrument, Sample, and Test Construction

The initial test form used for this study was a national assessment that is used for professional licensure purposes. The 119-item four-choice multiple-choice test is composed of four content categories. Each category contains about 30 items.

Two separate samples of examinees from different test administrations form populations P and Q. The difference in scores in the two samples/populations, as measured by this total test,

was about 0.27 of the (average) standard deviation on the test form. In both samples, the 119-item test form was split in order to construct two unique test forms, *X* and *Y*, and an anchor *A*. Test Forms *X* and *Y* were parallel in content but were intentionally designed to differ in difficulty, requiring one test to be equated to the other to place them on a common scale. The mean percent correct in the total sample *T* for *X* was 80.98, and for *Y* it was 61.71 (Table 2). The mean percent correct of the anchor test *A* was 69.53 (Table 3). In addition to summary statistics, differential item functioning (DIF) analyses (using the Mantel-Haenzsel criterion; Dorans & Holland, 1993) for gender and administration date were performed. No items were flagged for significant DIF. The items selected for Forms *X* and *Y* and for Anchor *A* are identified in the appendix.

Table 2

Summary Statistics for the Observed Frequencies of Test Forms X and Y in Populations P, Q, and T

	P (*N* = 6,168)		*Q* (*N* = 4,237)		*T* (*N* = 10,405)	
	X	*Y*	*X*	*Y*	*X*	*Y*
Mean	35.12	26.59	36.38	27.97	35.63	27.15
Mean (%)	79.82	60.40	82.68	63.57	80.98	61.71
SD	5.69	6.68	4.77	6.29	5.37	6.56
Skewness	–0.96	–0.10	–1.09	–0.27	–1.04	–0.18
Kurtosis	3.77	2.45	4.54	2.59	4.13	2.49
Obs. min	8	7	10	8	8	7
Obs. max	44	43	44	43	44	43
Alpha reliability	0.81	0.81	0.77	0.79	0.80	0.80

Equating Methods

Loglinear models were used to separately smooth the data for all equating methods; five moments of the univariate distributions (for the EG design) and the four moments of the marginals of the bivariate distributions (for the NEAT design) were preserved (Holland & Thayer, 2000). In the original investigation, examination of fit statistics provided evidence that there was a significant benefit in preserving four moments for the interaction rather than just one (von Davier et al., 2005).

Table 3

Sample Sizes, Means, Means as a Percentage of Total Score, Standard Deviations, and Alpha Reliabilities of the Scores on A_1, A_2, and A_3, in P, Q , and T

	P (*N* = 6,168)			*Q* (*N* = 4,237)			*T* (*N* = 10,405)		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
Mean	16.03	13.66	10.84	17.00	14.48	11.50	16.43	13.99	11.11
Mean (%)	66.79	68.30	67.75	70.83	72.40	71.88	68.46	69.95	69.44
SD	4.19	3.55	3.01	3.85	3.30	2.82	4.09	3.47	2.95
Alpha reliability	0.75	0.71	0.68	0.73	0.69	0.66	0.75	0.71	0.68

In the NEAT design, the following analyses were conducted for equating *X* scores to *Y* scores: equipercentile (frequency estimation) equating, chained equipercentile equating, KE PSE with optimal bandwidths that approximates the frequency estimation equating method, and KE PSE linear with large bandwidths that approximates the Braun and Holland (1982) linear equating method. In the EG design, which acted as the criterion, equipercentile (frequency estimation) equating was used. Given the differences in the shapes of the distributions between *X* in *P* and *Y* in *Q* that existed in our data, a nonlinear conversion is an appropriate choice as a criterion over one that is linear.

Because both Forms *X* and *Y* were created from one parent form, data existed for equating via both an EG and a NEAT design (See Table 1). For the EG design, data from both *P* and *Q* were used for equating. For the NEAT design, scores from *X* on *P* were equated to scores from *Y* on *Q*. Both *X* and *Y* contained 44 unique items. An additional set of items, *A*, which was substantively representative of *X* and *Y*, acted as an external anchor to *X* and *Y* for equating using the NEAT design. The length of *A* was varied to three sizes: (a) 24 items, (b) 20 items, and (c) 16 items.

Results

Summary Statistics

Because the original test form was split into two forms *X* and *Y*, data for *P* and *Q* existed for both test forms. Raw summary statistics for *X* and *Y* in both *P* and *Q*, are presented in Table 2. Overall, sample *Q* performed better on both test forms. *P* had mean scores of 35.12 (*SD* = 5.69) and 26.59 (*SD* = 6.68) for forms *X* and *Y* respectively, while *Q* had mean scores of 36.38 (*SD* =

4.77) and 27.97 (*SD* = 6.29) on forms *X* and *Y* respectively. In the target population *T*, *X* (*M* = 35.63, *SD* = 5.37) was less difficult than *Y* (*M* = 27.15, *SD* = 6.56) by 8.48 points (about one and one-half standard deviations). When expressed as percent correct, the mean raw scores were 79.82, 82.68, and 80.98 on *X* and 60.40, 63.57, and 61.71 on *Y* in *P*, *Q,* and *T*, respectively. The reliabilities of *X* and *Y* ranged from 0.77 to 0.81 across *P*, *Q,* and *T*.

The examinees sampled from *Q* also outperformed those from *P* on the anchor, and the scores were consistent across all anchor lengths (see Table 3). In *P*, the mean percent correct was 66.80, 68.28, and 67.76, while in *Q*, the mean percent correct was 70.85, 72.39, and 71.86 for the 24-, 20- and 16-item anchor tests, respectively. In *T*, the mean percent correct was 68.45, 69.95, and 69.43 for the 24-, 20- and 16-item anchor tests respectively. The correlations between each form and anchor test length were relatively high, ranging between 0.71 and 0.79 (see Table 4) in *P*, *Q,* and *T*. As expected, the correlations between the anchor and total test decreased as the number of items in the anchor test decreased from 24 to 16 items. Similarly, the reliabilities also decreased in *P*, *Q,* and *T* as the number of items in the anchor test decreased. In the 24-item anchor, reliabilities ranged from 0.73 to 0.75, while in the 16-item anchor reliabilities ranged from 0.66 to 0.68.

Table 4

Correlations Between Test Forms X, Y, and Anchor Tests A_1, A_2, and A_3 in Populations P, Q, and T

Correlation	*P* (*N* = 6,168)	*Q* (*N* = 4,237)	*T* (*N* = 10,405)
(X, A_1)	0.78	0.77	0.74
(X, A_2)	0.76	0.75	0.72
(X, A_3)	0.75	0.74	0.71
(Y, A_1)	0.79	0.78	0.76
(Y, A_2)	0.77	0.76	0.74
(Y, A_3)	0.76	0.75	0.73

Equating Conversions

Figures 1, 2, and 3 display the equating functions for the criterion EG equipercentile method and all NEAT equating methods for the 24-, 20- and 16-item anchor length conditions respectively. These graphs indicate very little difference in the results between the criterion and

nonlinear methods, equipercentile, chained equipercentile, and KE PSE (optimal bandwidth). The KE PSE (large bandwidth) method created a linear equating function, which was very similar to the other methods in the region of the mean test scores but differed from the other methods in other regions, particularly in the lower scores. The same pattern of results was observed across all external anchor lengths.

Differences Between Criterion Function and Method Function

Figure 4 illustrates the differences in performance of each equating method across the raw score scale when a 24-item external anchor was used. These results are representative of the relative performance of all of the methods across external anchor lengths. With a 24-item external anchor, the nonlinear methods—equipercentile, chained equipercentile, and KE PSE optimal bandwidths—all produced conversions that were very similar to the criterion, and once rounded, the scores would have been indiscernible from each other and from the criterion (i.e., smaller than DTM). The linear method KE large *h* did not meet the criterion well, and it produced differences much larger than the other methods across most of the scale.

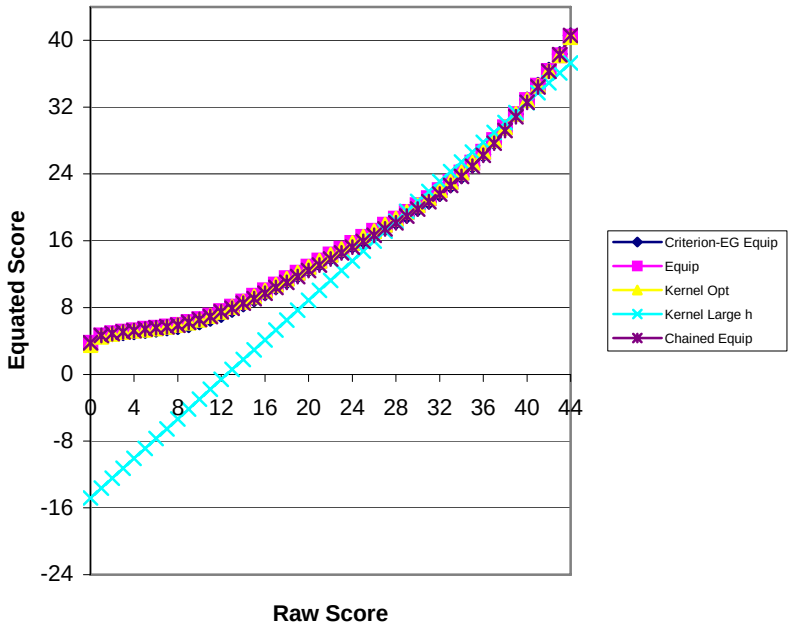

Figure 1. **Equating functions for criterion (EG equipercentile) and all other NEAT equating methods with 24-item external anchor.**

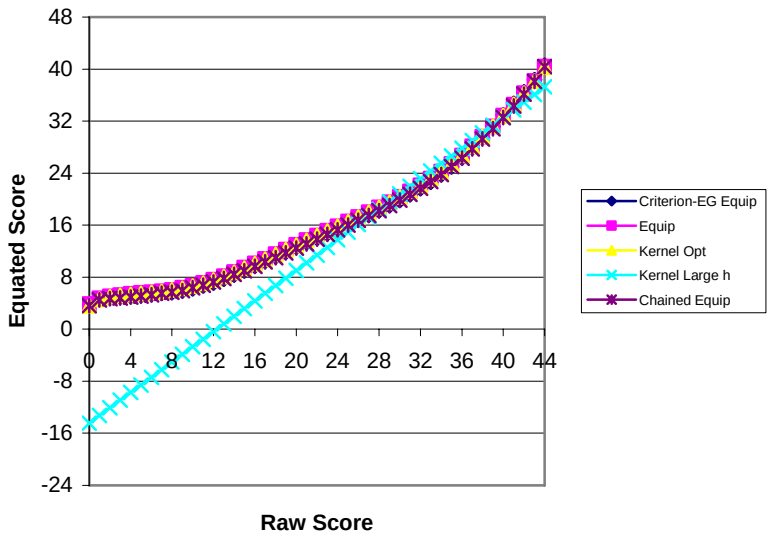

Figure 2. **Equating functions for criterion (EG equipercentile) and all other NEAT equating methods with 20-item external anchor.**

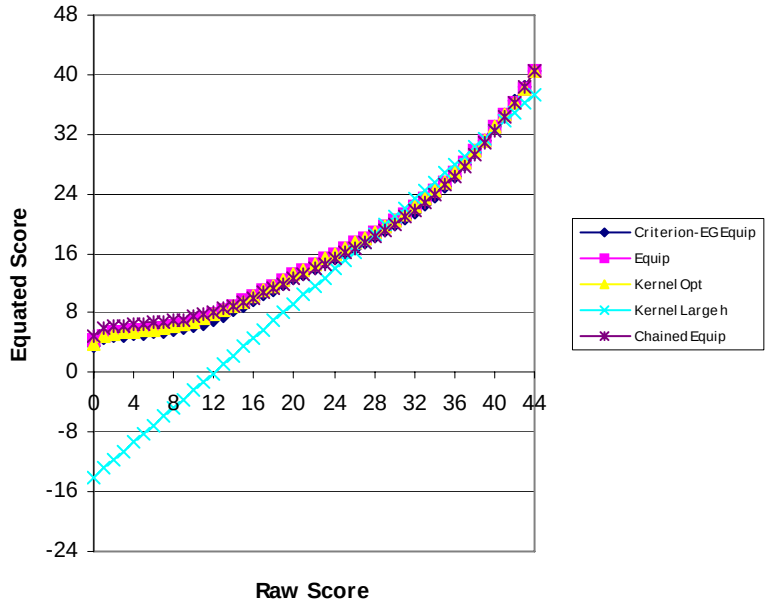

Figure 3. **Equating functions for Criterion (EG equipercentile) and all other NEAT equating methods with 16-item external anchor.**

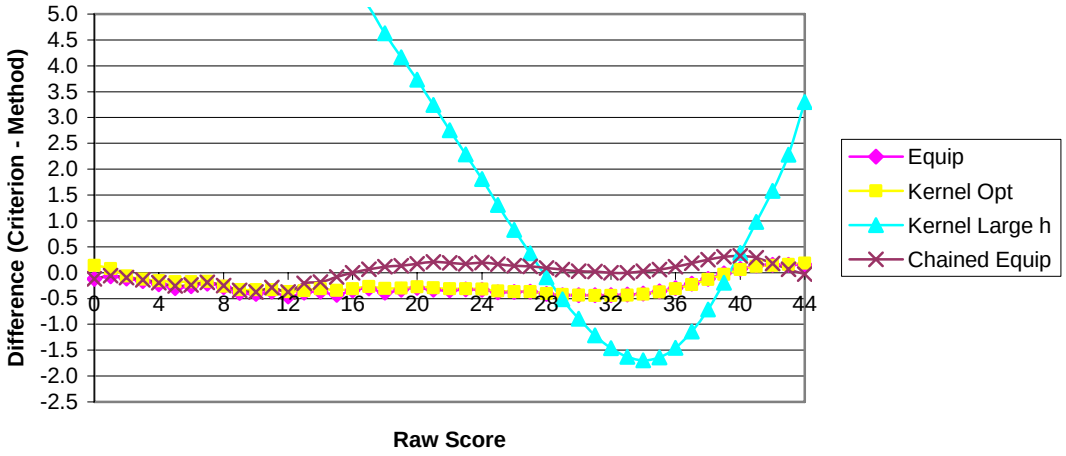

Figure 4. **Conditional difference at each raw scale score point (relative to EG equipercentile criterion) of equipercentile, KE PSE with optimal *h*, KE PSE with large *h*, and chained equipercentile equating methods as a function of raw score with a 24-item external anchor.**

Figures 5–8 depict the performance of each method relative to the criterion across external anchor test lengths. Figure 5 shows the differences for KE PSE with optimal bandwidths. As the anchor test length decreased, the conversion became more different from the criterion across the entire score scale. At 24 and 20 items, the differences were smaller than a DTM, with the exception of score points 29–31 with a 20-item external anchor. With 16 items, score differences were larger and would therefore be observable except at the extreme low and high ends of the score scale.

Figure 6 shows the differences for equipercentile (frequency estimation) equating. The results were very similar to KE PSE optimal, but were larger than a DTM at score points 28-34 with a 20-item external anchor. With a 16-item external anchor, the differences were larger than DTM except at the high end of the score scale. Figure 7 shows the chained equipercentile equating differences. When the anchor test length was 24 or 20 items, the differences were smaller than a DTM across the entire score scale. With a 16-item anchor, the differences were also smaller than a DTM, except from points 0-15 on the raw score scale, where the differences observed were larger than those observed in either equipercentile or KE PSE optimal equating. Figure 8 shows the differences for KE PSE large *h*. This method had detectable score differences (relative to the criterion equating function) for all anchor test lengths across the entire score scale.

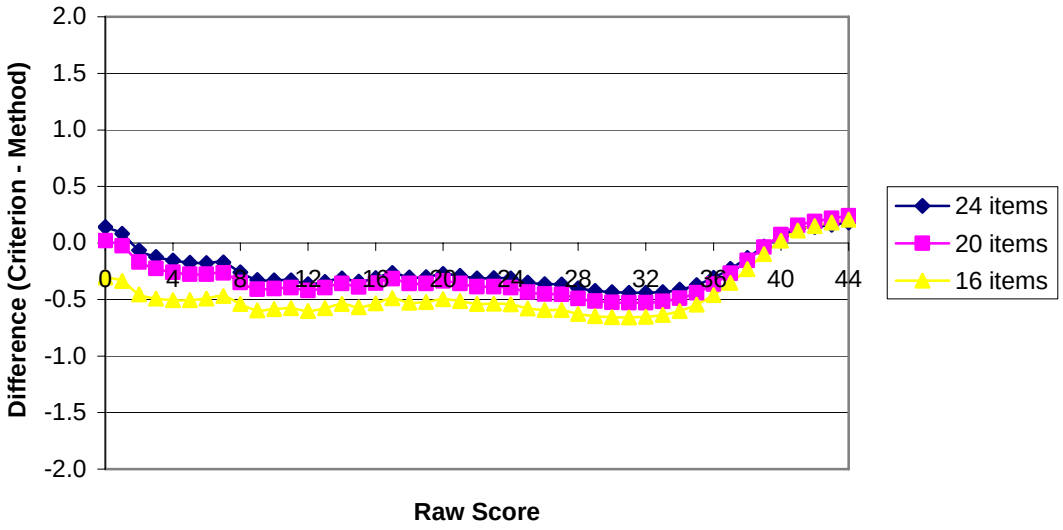

Figure 5. **Conditional difference at each raw scale point (relative to EG equipercentile criterion) of NEAT KE PSE with optimal bandwidths as a function of raw score for 24-, 20-, and 16-item external anchor test lengths.**

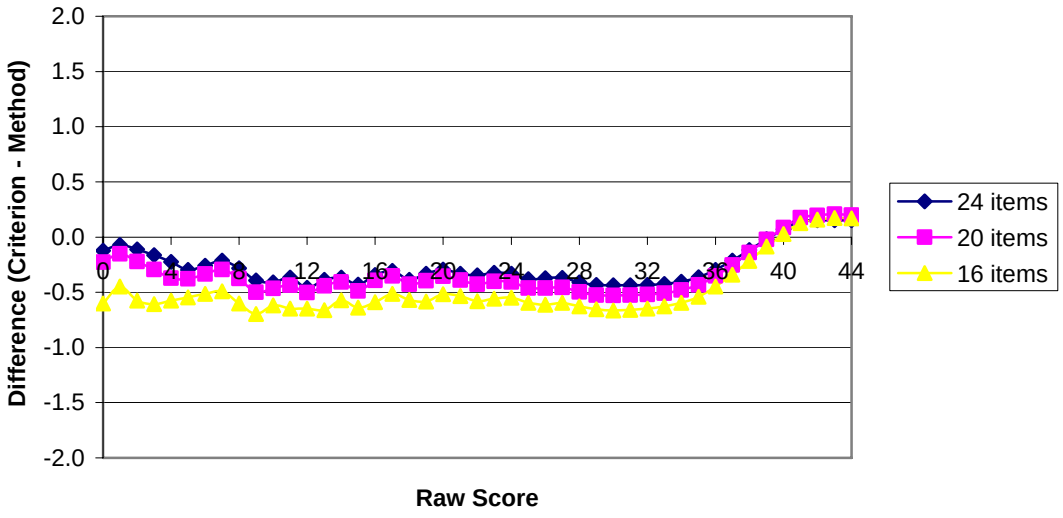

Figure 6. **Conditional difference at each raw scale score point (relative to EG equipercentile criterion) of NEAT equipercentile (frequency estimation) equating as a function of raw score for 24-, 20-, and 16-item external anchor test lengths.**

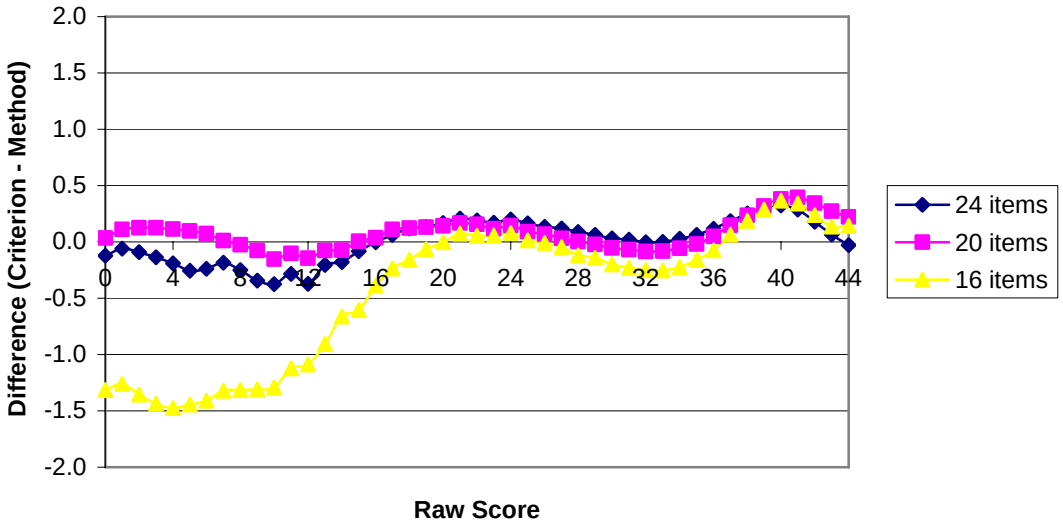

Figure 7. **Conditional difference at each raw scale score point (relative to EG equipercentile criterion) of NEAT chained equipercentile equating as a function of raw score for 24-, 20-, and 16-item external anchor test lengths.**

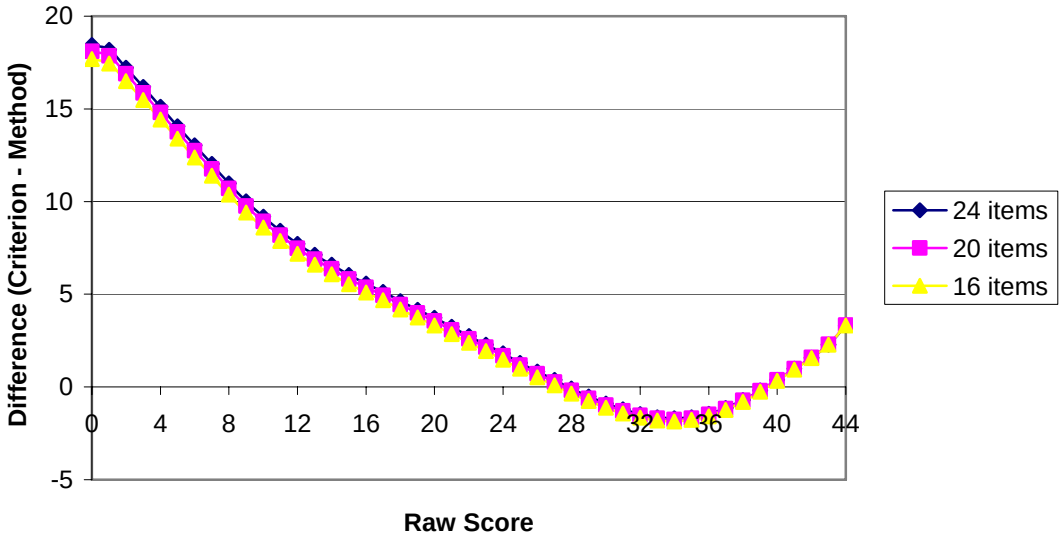

Figure 8. **Conditional difference at each raw scale score point (relative to EG equipercentile criterion) of NEAT KE PSE with large *h* as a function of raw score for 24-, 20-, and 16-item external anchor test lengths.**

Root Mean Square Difference

No large differences in any of the equating functions were observed across anchor test lengths. $RMSE = \sqrt{\bar{d}^2 + sd_d^2}$, where $\bar{d}$ is the mean of the conditional equating differences for each NEAT equating method with the criterion and *sd* is the standard deviations of these differences.

Measures of RMSE indicate that the equipercentile, chained equipercentile, and KE PSE (optimal bandwidth) methods produced small errors in comparison to the KE PSE (large bandwidth) method (see Table 5). Comparing each method across anchor test lengths, the chained equipercentile method produced the smallest RMSE at 24- and 20-item anchor lengths (0.19 and 0.16, respectively, but when the anchor length was reduced to 16 items, the error increased dramatically, and it had the largest RMSE (0.76). The equipercentile frequency estimation (0.32, 0.39, and 0.54 for 24, 20, and 16 items, respectively) and KE PSE with optimal bandwidths (0.30, 0.36, and 0.51 for 24, 20, and 16 items, respectively) also had an increased RMSE as the anchor test length decreased, though to a much lesser extent. The RMSE for KE PSE (large bandwidth) remained relatively constant across anchor test lengths, but was much larger than that for the other, nonlinear equating methods (7.82, 7.64, and 7.43 for 24, 20, and 16 items, respectively).

Table 5

Root Mean Square Difference (Error; RMSE) as a Function of Method and Anchor Test Length Versus the EG Equipercentile Criterion

Method	A_1	A_2	A_3
	24 items	20 items	16 items
Equipercentile	0.32	0.39	0.54
Chained equipercentile	0.19	0.16	0.76
KE PSE—optimal	0.30	0.36	0.51
KE PSE-large *h*	7.82	7.64	7.43

Note. $RMSE = \sqrt{\bar{d}^2 + sd_d^2}$

Discussion

Anchor Test Length

In general, the equating bias associated with each method increased, but not dramatically so, as the anchor test length decreased, with the exception of the chained equipercentile method. The bias associated with this method increased as the number of anchor items dropped from 20 to 16. This result is not surprising, given that this procedure chains the forms together via equating each form to the common items, making anchor length a more important factor than for the other methods (Kolen & Brennan, 2004). The linear KE PSE (large bandwidth) method produced the largest errors when compared to the criterion, especially in RMSE, which was also expected given that it produces a linear equating function while all of the other methods (including the criterion) are nonlinear. The results for this method would likely look more promising if compared against a linear criterion.

The modest changes in the equating robustness of the equipercentile frequency estimation and both KE PSE methods, which rely on relatively stable correlations between the test forms and anchor forms as the anchor length decreases, are likely attributable to excellent test construction. These methods rely on the invariance of the conditional distributions of the forms *X* and *Y* on the anchor *A* across *P* and *Q*. Budescu (1985) suggested that the correlation was the most important factor in managing equating error. Another related factor is that the reliabilities of *A* did not decrease a great deal across different anchor test lengths (Table 3).

The KE PSE with large *h* method had an observed RMSE that was much larger than those observed for the other methods. This result occurred largely because this method is linear, while the other methods and the criterion equating method are nonlinear. The KE PSE with large *h* does a poor job of aligning scores in the lower end of the score scale where there were no data, but was much closer, though still not as accurate, in the region where most of the data were present (raw scores of approximately 27-41). Given that *X* and *Y* differed in distribution shape, using a linear conversion was not appropriate.

Influence of Criterion Selection

Whenever a criterion is selected, the choice of criterion will ultimately influence the results. The EG/combined group method was chosen to preserve the common population for which the conversion holds. The equipercentile method was chosen because the two forms, *X*

and *Y*, differed in means, variances, and skewness. The polynomial loglinear model that fits the first five moments for each univariate distribution was chosen based on various fit statistics.

The results of this study suggest that the results using KE PSE with optimal bandwidths, equipercentile frequency estimation, and chained equipercentile in a NEAT design produced very similar results to the EG equipercentile equating, particularly at the high and low ends of the score scale, with 24- and 20-item anchors. The interpretation of this result must be made carefully, because a cursory look might conclude that the NEAT equating methods were most accurate in the low and high regions of the score scale. In reality, the criterion EG equipercentile would be expected to produce an extremely accurate conversion in the region where most of the data were observed, with less accuracy at high and low scores, because there are fewer cases in those regions.

Conclusions

Practitioners are frequently faced with choosing the best equating method for a particular application when using a NEAT design, without the benefit of having an EG criterion to help guide the decision-making process. Practitioners might also face a trade-off between maximizing the anchor length for statistical purposes and minimizing it for other considerations, including test security and item datedness. These results suggest that the choice of equating method can change the amount of error present in the test scores, particularly with shorter anchor lengths (in this case, 16 items). Overall, the KE PSE with optimal bandwidths method performed comparably with classical and chained equipercentile methods in the NEAT design when the EG equipercentile method was used as a criterion. On the other hand, KE PSE with large bandwidths performed poorly when compared to the nonlinear criterion.

References

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

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22,* 13–20.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225–244.

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.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

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

Ricker, K. L., & Gierl, M. J. (2005, April). *The consequences of multidimensionality to IRT equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

von Davier, A. A., & Han, N. (2004, April). *Population invariance and linear equating for non-equivalent groups design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2005, April). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed 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. (2004). *The kernel method of test equating*. New York: Springer-Verlag.

Appendix

Test Form Construction for Forms X, Y, and Anchor Test A

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