A Comparative Evaluation of Kernel Equating and Test Characteristic Curve Equating

Applied Psychological Measurement

2018, Vol. 42(2) 155–168

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617712245

journals.sagepub.com/home/apm

R. J. De Ayala¹, Bradley Smith¹, and Rebecca Norman Dvorak²

Abstract

This study compares the kernel equating (KE) and test characteristic curve (TCC) equating methods using the nonequivalent anchor test equating design. In this Monte Carlo study, four independent variables were examined: sample size, test length, average form discrimination, anchor test reliability, and the percentage of anchor items. For each condition, there were 100 replications. To assess the performance of TCC equating and KE, the differences between the examinee parametric true scores and the equated estimated expected true scores were examined. The equated scores were based on the average across replications for each condition. Generally speaking, both KE and TCC equating produced accurate results, although KE tended to perform better than TCC on the parametric true score scale across conditions. Past research and the current study's results seem to indicate that KE should be strongly considered for most equating situations, particularly in light of its flexibility.

Keywords

equating, test characteristic curve, kernel equating

In large-scale educational assessments (e.g., district- and statewide assessments, college admissions tests, professional licensure tests), there is often the need to administer assessments at multiple time points. For example, in college admission tests students are often provided the opportunity to choose when they would like to be tested (Kolen & Brennan, 2004). As a result, to address testing integrity and security there is a need to generate multiple alternate test forms with each form administered at one of the time points. Despite a test developer's attempt to create alternate forms that are equal in difficulty, most likely the forms' difficulties will differ, albeit slightly. To adjust for these differences, test equating methods have been developed. As a result, these methods allow the scores earned by examinees on alternate test forms to be directly compared by placing test scores from alternate forms on the same scale. Examples of equating methods include mean, linear, equipercentile, kernel equating (KE; Holland & Thayer, 1989), and item response theory (IRT) equating techniques such as the test characteristic curve (TCC; Stocking & Lord, 1983) method.

¹University of Nebraska–Lincoln, NE, USA

²Human Resources Research Organization, Alexandria, VA, USA

Corresponding Author:

R. J. De Ayala, University of Nebraska–Lincoln, TEAC 114, Lincoln, NE 68588, USA

Email: rdeayala2@unl.edu

Because significant educational and high-stakes decisions (e.g., school funding, admission, certification, graduation) are often made about examinees or schools based on equating results, it is important to understand how choosing one method over another may affect decisions. In this study, the comparative performance of the TCC and KE methods was examined.

TCC Equating

The Stocking and Lord (1983) TCC is a parametric IRT equating method of true score equating that determines the equating coefficients based on TCCs (Baker, Al-Karni, 1991). As such, it utilizes information from all item parameter estimates to derive the equating coefficients. In addition to its frequent use in practice, this method has been shown to be robust to the distributional characteristics of the item parameter estimates (Baker, Al-Karni, 1991; Cook & Petersen, 1987; Hanson & Beguin, 2002; Kaskowitz & De Ayala, 2001; Kolen & Brennan, 2004; Stocking and Lord, 1983).

The gist of the TCC method is to minimize the differences between the Form X and Form Y TCCs. The TCC specifies the relationship between the expected true score, τ , on a test and the latent ability, where τ for examinee j is given by $\tau_j \equiv \tau(\theta_j) \equiv \sum_{i=1}^m p_{ij}(\theta_j; \vec{v})$, m is the number of test items, θ_j is examinee j 's ability, and $\vec{v}$ is a vector containing the model's item parameters. The TCC method requires the use of common items (anchor test) that are administered to those examinees administered Form X and those administered Form Y .

In TCC equating, the estimation of the equating coefficients requires first determining the TCC for each form, followed by minimizing a quadratic loss function defined in terms of both TCCs. Specifically, the method seeks to minimize the square of the differences between Form X 's and Form Y 's TCCs (Stocking & Lord, 1983): $SqrDiff_j(\theta) = \left[\sum_{i:A} p_{ij}(\hat{\theta}_j; \vec{v}_{Xi}) - \sum_{i:A} p_{ij}(\hat{\theta}_j; \vec{v}_{Yi}^*) \right]^2$, where $\hat{\theta}_j$ is examinee j 's estimated ability (cf. Baker, Al-Karni, 1991; Stocking & Lord, 1983); A indicates items on the anchor test. For the 2PL model, $\vec{v}_{Xi}$ contains Form X 's item i 's estimated discrimination ($\hat{a}_{Xi}$) and difficulty parameters ($\hat{b}_{Xi}$; that is, $(\vec{v}_{Xi})' = [\hat{a}_{Xi}, \hat{b}_{Xi}]$). Furthermore, $\vec{v}_{Yi}^*$ contains Form Y item i 's discrimination and difficulty parameter estimates linearly transformed to Form X 's metric, $\hat{a}_{Yi}^*$ and $\hat{b}_{Yi}^*$, respectively (i.e., $\hat{a}_{Yi}^* = \hat{a}_{Yi}/\hat{\alpha}$, $\hat{b}_{Yi}^* = \hat{\alpha}\hat{b}_{Yi} + \hat{\kappa}$; $(\vec{v}_{Yi}^*)' = [\hat{a}_{Yi}^*, \hat{b}_{Yi}^*]$). As can be seen, $SqrDiff_j(\theta)$ is the squared difference between Form Y 's and X 's TCCs for a given ability parameter for item i . When accumulated over the ability scale, this term forms the loss function, $F = 1/N \sum_j SqrDiff_j(\theta)$, whose minimization yields the estimated values of the equating coefficients α and κ ; N represents total number of examinees or the number of arbitrary equidistant ability scale points. Thus, provisional estimates of $\hat{\alpha}$ and $\hat{\kappa}$ are used to transform $\hat{a}_{Yi}$ and $\hat{b}_{Yi}$ to be on Form X 's (i.e., target or reference) metric ($\hat{a}_{Yi}^*$ and $\hat{b}_{Yi}^*$, respectively). In the next iteration, these provisional estimates of $\hat{\alpha}$ and $\hat{\kappa}$ are refined to further minimize the discrepancy between the two characteristic curves, and improved $\hat{a}_{Yi}^*$ and $\hat{b}_{Yi}^*$ are obtained. This iterative process is repeated until the difference between estimates from successive iterations is less than the convergence criterion or the maximum number of iterations is reached.

The equating coefficients α and κ represent the slope and intercept, respectively, of the linear equation to transform the metric of one test form's calibration to that of the other. Once the equating coefficients are determined, then the person ability estimates on the Y metric are transformed to the X metric by $\hat{\theta}_{Yj} = \hat{\alpha}\hat{\theta}_{Yj} + \hat{\kappa}$. Because TCC equating is based on IRT, the model's relevant assumptions (e.g., dimensionality, local independence, functional form) need to be

tenable. If the assumptions of parametric IRT are not sufficiently met in practice, then one may have inaccurate estimation of the equating coefficients.

KE

When the assumptions for a parametric method are untenable, then a nonparametric approach may be beneficial. KE is a nonparametric method described by von Davier, Holland, and Thayer (2004b) as “a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contains the linear equating function as a special case” (p. 45). In this regard, fewer mathematical assumptions are made by using this kernel than with linear interpolation (von Davier et al., 2004b).

There are two mathematical methods that can be used for KE. In chain equating (CE), one first links Form X to the anchor test A , followed by linking A to Form Y . In contrast, with post-stratification equating (PSE), the marginal distributions of both X and Y in the target population, T , are estimated first, and then the equating function is computed (von Davier, Holland, & Thayer, 2004a). T is a mixture of the two populations from which the examinee groups taking forms X or Y arise.

The general process of performing KE involves five steps: (a) presmoothing (if necessary), (b) estimation of the score probabilities (i.e., the probabilities of the possible discrete values), (c) continuization, (d) equating, and (e) calculating the standard error of equating. The presmoothing step involves estimating the score probabilities for a particular equating design by fitting a statistical model (e.g., a log-linear model, IRT model) to each of the form–anchor test score distributions. The resulting smoothed score distributions are used for the remaining equating process as well as providing the matrix that can be used for calculating the standard error of equating (Chen, 2012; von Davier et al., 2006).

In the second step, the score probabilities are estimated. PSE and CE differ in this step. In PSE, A is used to stratify both X and Y by conditioning their distributions on A , and then reweighting their conditional distributions to estimate the score distributions for Forms X and Y in T ; each one of these conditional distributions is assumed to be population invariant (von Davier et al., 2004a). In contrast, for CE the hypothetical target population does not play an explicit role and does not appear in the composed link from X to A to Y (von Davier et al., 2004a). This composed link function reflects the combining of two single-group designs (i.e., X to A , A to Y) to obtain the equating function, and thereby avoids the need to first estimate the score probabilities (von Davier et al., 2004a; von Davier et al., 2004b). In addition, each one of these links is assumed to be population invariant.

The third step is the continuization step, and involves the process of transforming the discrete score distributions for X and Y on the target population into continuous score distributions over the entire score range by using kernel smoothing (von Davier et al., 2006). It is this step that involves the Gaussian kernel smoothing of the distributions of X and Y as well as the continuization constants (bandwidths). In effect, this step consists of selecting these continuization constants. The bandwidth can be thought of as the width of a sector that traverses the distribution transforming the data within each sector. This transformation is a function of which kernel density estimator is used (e.g., a Gaussian kernel). As such, the bandwidth serves as a smoothing parameter with values close to 0 yielding very little smoothing and increasing values resulting in correspondingly increased degrees of smoothing. Thus, the bandwidth affects the shape of the resulting continuous approximation (Holland & Thayer, 1989). A form's bandwidth value can be specified by the researcher or algorithmically determined to minimize the mean square error (*MSE*).

TCC and KE Studies

There have been only a few studies that directly compare KE with traditional methods. These studies have been based on empirical test data, and have provided evidence that KE is a viable alternative to traditional methods (e.g., Mao et al., 2006; von Davier et al., 2004b). Specifically, KE appears to perform very similar to, and in some circumstances better than, traditional linear and equipercentile methods (Mao et al., 2006; von Davier et al., 2004b; von Davier et al., 2006). Other applications of kernel smoothing have demonstrated equivalent performance to parametric continuous response IRT methods (Ferrando, 2004). However, previous research also suggests that KE is affected by some of the same factors that affect IRT, such as sample size and test length (Lee, 2007). These factors include the number of anchor items, the sample size, and the test length. Increasing the number of anchor items in IRT equating has been shown to reduce MSEs. Moreover, larger sample sizes are also associated with lower MSEs in IRT (Hanson & Beguin, 2002). Similarly, in KE the fit of kernel smoothing has been found to improve with increased sample sizes (Lee, 2007). This research also suggests that kernel smoothing improves with increased test length. Previous research involving KE and TCC equating has found that KE(CE) performed well although the TCC method produced more favorable results (Godfrey, 2007). Furthermore, Meng (2012) found that TCC's comparative accuracy (with respect to KE(PSE) and mean/sigma) was dependent on the score scale region where the comparisons were made. This study's objective was to extend the KE–TCC equating literature by assessing the comparative equating accuracy of TCC, CE, and PSE. Utilizing a replication design and a true score criterion, the effects of differential (a) item discrimination, (b) anchor test reliability, (c) sample size, (d) test length, and (e) anchor test length were examined.

Method

To compare the equating accuracy of KE and TCC, a Monte Carlo study was conducted with two forms using a common item external anchor test (i.e., a NEAT equating design). In the following, the independent variables, the data generation process, the equating procedure, and the dependent variables are presented.

Independent Variables

Based on the literature as well as what is generally found in educational testing situations, it was expected that certain variables would affect the accuracy of KE and TCC equating. These variables are the sample size, the test length, the average form discrimination, and the percentage of anchor items. In addition, Ricker and von Davier (2007), in a study on the impact of external anchor test length, found that bias tended to increase as the anchor test length decreased.

Test length. This factor had three levels: 25, 50, and 100 items. Equating studies conducted by Educational Testing Services (ETS) tend to use test lengths of close to or more than 100 items (Mao et al., 2006; von Davier et al., 2006). Therefore, the 100-item level may be seen as the low end of a long educational test. Conversely, a “short test” such as a 25-item test might be used in a high school graduation demonstration exam, such as in mathematics.

Percentage of anchor items. The number of anchor items was implemented as a percentage of test length. There were three levels (10%, 20%, 30%) for the 50- and 100-item test lengths. For the 25-item test length, there were two levels (20%, 30%) because a 10% level would yield a three-item anchor test which, in practice, is highly improbable. For all test lengths, the shorter anchor test was subsumed by the next longer anchor test (e.g., the 10% anchor condition was a subset

of the 20% condition). It is generally accepted in common item equating that at least 20% of the items on an assessment should be anchor items, although a minimum of 15 items has also been suggested (Kolen & Brennan, 2004). However, studies have not specified the maximum number of anchor items as a function of test length (Sinharay & Holland, 2007).

Average form discrimination. This factor's levels were low, medium, and high; details are provided below.

Sample size. Three levels of sample size were investigated: 1,000, 2,000, and 10,000 simulees with the 1,000 and 2,000 conditions reflecting likely sample sizes at the school district level with the 10,000 level being a realistic size for a statewide assessment.

Anchor test reliability. To examine whether the effect of anchor test length is due to an increase (or decrease) in reliability (i.e., the longer a test the larger the reliability, all things being equal), anchor test reliability was examined. The reliability of the anchor test was allowed to vary or kept constant across the varying anchor test lengths conditional on test length (i.e., percentage of anchor items).

The crossing of these factors produced 144 cells. For each cell, 100 replications were conducted.

Data Generation

Data generation was a three-step process with the anchor tests created first, the test forms created second, and the response data generated third. For random number generation, different seeds were used for each form, each stratum, each anchor test length, each sample size, and each replication. Below, the anchor tests are discussed first followed by the test forms and response generation.

To create the anchor tests, item discrimination parameters were randomly generated from a lognormal distribution (e.g., $M = 0$, $SD = 0.3$). In the constant reliability condition, it was necessary to modify these item discriminations to approximate a target average reliability of .90. (Online Appendix A contains the specifics of creating the constant reliability condition.) The overall average discrimination across the different anchors was 0.93 (overall $SD = 0.76$). For the nonconstant reliability condition, item discrimination parameters were randomly generated from a lognormal distribution and were modified as necessary to maintain approximately the same minimum and maximum discriminations. The overall average discrimination across the different anchors was 1.09 (overall $SD = 0.43$).

The θ continuum was divided into five strata ($-2.5 \dots -1.5$, $-1.5 \dots -0.5$, $-0.5 \dots 0.5$, $0.5 \dots 1.5$, $1.5 \dots 2.5$) to ensure that the anchor tests contained items that spanned the continuum. The midpoint of each stratum (i.e., -2 , -1 , 0 , 1 , 2) served as the mean for generating random normal deviates from a normal distribution ($SD = 0.25$). These random normal deviates served as the item difficulty parameters.

To generate the test forms X and Y with three levels of average discrimination, 100 items were randomly sampled from lognormal distributions, so that the resulting parameters varied in average discrimination but had similar variability. For each level of discrimination, the three test lengths were created by selecting the first 25 and 50 items from the 100-item set for a given discrimination level to serve as the item discriminations for the 25- and 50-item test lengths, respectively.

The test forms item difficulties were obtained by randomly generating 100 standard normal deviates for each form. Similar to what was done with the item discriminations, the first 25 and 50 items from the 100-item set for each form served as the item difficulties for the 25- and 50-

item test lengths. In those cases where the correlation between item discrimination and difficulty was negative (i.e., easy items that were highly discriminating and vice versa), the item parameters were sorted to obtain a nonnegative correlation and a more realistic item set.

Person abilities (θ) were randomly sampled from a $N(0,1)$ for each of the three different sample sizes (1,000; 2,000; 10,000). In addition to the simulee's θ , the examinee's true score was calculated. Because this true score uses person and item parameters, it is referred to as the *parametric* expected true score ("parametric true score" for short). Using the appropriate item parameters and the simulee's ability, the probability of a correct response to an item was calculated according to the 2PL model. This probability was compared with a randomly generated uniform number $[0,1]$. If the probability was equal to or greater than the uniform random number, then the simulee's response was coded 1, 0 otherwise. This process was repeated for all items and for all simulees.

TCC Equating Procedure

To perform TCC equating, the data were calibrated using the 2PL model and BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003) utilizing 100 quadrature points for both item calibration and person estimation; EM and Newton maxima were set at 100. The 2PL model was selected because this model has commonly been chosen in previous equating simulation studies (Ferrando, 2004; Lee, 2007; Sinharay & Holland, 2007). Expected a posteriori (EAP) was used for person estimation.

Subsequent to the calibration, the Form Y item parameter estimates were linked to Form X 's metric using the program EQUATE 2.1 (Baker, Al-Karni, Al-Dosary, 1991) with 25 evaluation points (i.e., $N = 25$) and the maximum number of iterations set to 100. In addition, EQUATE was used to equate the Form Y $\hat{\theta}$ s to the Form X metric. These person ability estimates were then transformed into their corresponding *estimated* expected true scores ("estimated true scores" for short) 1 and the item parameter estimates. As such, each simulee had a parametric expected true score based on θ ("parametric true score") and one calculated using the simulee's $\hat{\theta}$, the estimated expected true score ("estimated true score").

KE Procedure

KE was performed using the R package kequate (Andersson, Bränberg, & Wiberg, 2013). This package consists of three phases involving the creation of frequency distributions (*kefreq*), fitting a generalized linear model via the *glm* function, and performing the equating with *kequate*. For *glm*, the maximum iteration was set to 100, and the Poisson distribution was used. The optimal bandwidth was selected by the program. For each replication, 100 models varying in complexity (i.e., from nonadditive to additive models as well as with varying powers) were fitted. The model selected (i.e., the modal model) for equating had the lowest Akaike's Information Criterion (AIC) the most times across the 100 replications within a cell across Forms X and Y . Typically, the modal model had the form of $X, X^2, X^3, X^4, A, A^2, A^3, XA$ or $X, X^2, X^3, X^4, A, A^2, A^3, A^4, XA$, where X represents a test form, A represents the anchor test, and XA is a first-order interaction. Once the best model was identified, it was applied to each data set for the equating of Form Y scores with Form X scores using PSE and CE.

Dependent Variables

To assess the performance of TCC equating and KE, the differences between the simulee ability estimates were compared. Because KE results are on a number correct scale, the TCC equated

Form Y to Form X ability estimates ($\hat{\theta}_{Y|X}$ s) were transformed to their corresponding estimated expected true scores ($\hat{\tau}_{TCC}$). For KE, the CE and PSE equated observed scores, $\hat{\tau}_{KE-CE}$ and $\hat{\tau}_{KE-PSE}$, respectively, were used as the simulee ability estimates.

The true scores are on a metric defined by the number of questions correctly answered and have a maximum given by the test length. Thus, to facilitate comparisons across instruments of different lengths, root mean square error ($RMSE$) and bias were calculated using the proportion correct scores (domain scores) of the estimated and parametric expected true scores.

$RMSE_{\pi}$ and $Bias_{\pi}$ were calculated for each method pairing to determine which approach, overall, best approximated the simulees' parametric true scores. $RMSE_{\pi}$ was given as follows:

$$RMSE_{\pi} = \sqrt{\frac{\sum (\hat{\pi} - \pi_P)^2}{K}}, \quad (1)$$

where $\hat{\pi}$ is $\hat{\tau}_{TCC}/m$, $\hat{\tau}_{KE-CE}/m$, or $\hat{\tau}_{KE-PSE}/m$; π_P is the domain parametric true score (τ_P/m); and K is the number of replications. Recall, $\hat{\tau}_{TCC}$ was obtained by using the linked Form Y item parameter estimates ($\hat{a}_{j(Y|X)}$ and $\hat{b}_{j(Y|X)}$) and the equated $\hat{\theta}_{Y|X}$, whereas τ_P utilized θ_Y and Form X 's item parameters, $a_{j(X)}$ and $b_{j(X)}$. Thus, τ_P is the expected true score for a Form Y person taking Form X using its item parameters (i.e., there is no estimation error). In addition, the bias of the person estimate was calculated as follows:

$$Bias_{\pi} = \frac{\sum (\hat{\pi} - \pi_P)}{K}, \quad (2)$$

where all terms are defined as above.

Results

Results showed that the mean discrimination is roughly equivalent across test lengths regardless of the average discrimination level and varies as one would expect across the discrimination-level factor. In addition, anchor test item difficulties tended, generally speaking, to fall between -2.39 and 2.46 with item discriminations for the nonconstant reliability condition averaging between 1.02 and 1.23 . In contrast, the mean item discrimination varied considerably in the constant reliability condition to maintain a relatively constant reliability across percent anchor lengths. With respect to estimated reliability estimates, the anchor test coefficient alphas for the constant reliability condition were similar across percent anchor lengths, test forms, and sample sizes with average coefficients ranging from 0.56 to 0.63 . In contrast, in the nonconstant reliability condition the average coefficient range showed the expected increase in reliability estimates as percent anchor increased and ranged from 0.32 to 0.81 . The average coefficient alphas for a given test length were similar across forms and sample sizes, and showed the expected progressive increase as test length increased. Detailed results are presented in Online Appendix B.

Equating Estimates

BILOG, EQUATE, and *glm* produced converged solutions for each replication in all conditions. In the following, the estimated true scores from each of the methods are referred to as the TCC, KE (CE), and KE (PSE) equated true scores.

As mentioned above, the bandwidths (h_X , h_Y) in KE control the amount of smoothing that occurs. Bandwidth descriptive statistics are found in Online Appendix B. The difference

between the average h_X and h_Y was no more than 0.0250 for KE (CE) and 0.0224 for KE (PSE), although, on average, h_Y was slightly larger than h_X . The average bandwidths varied from 0.4955 to 0.7176 with an overall mean bandwidth of 0.6248. As such, the equating functions used for KE were essentially reflecting equipercentile equating (see Häggström & Wiberg, 2014; von Davier et al., 2004b).

Generally speaking, the average $RMSE_\pi$ decreased as test length increased within percent anchor level. For example, for the sample size of 1,000 as test length increased from 25 to 50 to 100 items average $RMSE_\pi$ decreased from 0.088 to 0.048 to 0.035 for TCC for the 20% anchor level; similar results were found for KE (CE) and KE (PSE). This was true regardless of average discrimination level. KE (CE) and KE (PSE) tended to have very similar average $RMSE_\pi$ s regardless of test length, percent anchor, and average discrimination level, although KE (CE) exhibited slightly less overall bias. Except for a couple of conditions involving the medium or low discrimination levels and the 50-item test length conditions in which TCC had slightly smaller average $RMSE_\pi$ values, the KE approaches tended to slightly outperform TCC. In addition, there were very small differences (less than |0.001|) between the average $RMSE_\pi$ for the constant and nonconstant conditions; this was true regardless of sample size. When these differences occurred, they never involved TCC. The average $RMSE_\pi$ values for the 2,000 and 10,000 sample size conditions were extremely similar to those of the 1,000 condition. Comparisons of the conditions between the 1,000 and 2,000 sample size never exceeded |0.003|, |0.004| for the comparison between the 1,000/10,000 levels, and |0.003| for the comparison between the 2,000/10,000 levels. Average $RMSE_\pi$ values for the 1,000 level are presented in Online Appendix B; results for sample sizes of 2,000 and 10,000 were similar to those of the 1,000 level.

With respect to the average $Bias_\pi$ for KE (CE) and KE (PSE) tended to be less than that of TCC, although the average $Bias_\pi$ never exceeded the |0.062| value observed for TCC with the 25-item test length. In general, the average $Bias_\pi$ tended toward being slightly negative with exceptions to this pattern typically being associated with TCC. As was the case with $RMSE_\pi$, there were very small differences (less than |0.003|) between the average $Bias_\pi$ for the constant and nonconstant conditions; this was true irrespective of sample size. When these differences occurred, they never involved TCC. The average $Bias_\pi$ values for the 2,000 and 10,000 sample size conditions were similar to those of the 1,000 condition. The absolute differences between the 1,000 and 2,000 sample size conditions never exceeded 0.005, and for the 1,000 and 10,000 levels the differences ranged from -0.004 to 0.005. A comparison between the average $Bias_\pi$ differences for the 2,000 and 10,000 levels showed differences ranging from -0.002 to 0.003. Average $Bias_\pi$ values for the 1,000 level are presented in Online Appendix B; results for sample sizes of 2,000 and 10,000 levels were similar to those of the 1,000 level.

Although the average $RMSE_\pi$ and $Bias_\pi$ provide an overall performance picture for TCC, KE (CE), and KE (PSE), they do not provide an assessment of accuracy at different proficiency locations. To this end, $RMSE_\pi$ and $Bias_\pi$ were plotted as a function of domain score.

Examination of the $RMSE_\pi$ plots showed that, for a given test length, anchor length did not meaningfully affect the $RMSE_\pi$ s for any of the methods. Moreover, comparing the constant and nonconstant reliability levels' $RMSE_\pi$ for each cell showed extremely similar performance patterns. As a result, in the following only the nonconstant results are presented. With respect to test length, the $RMSE_\pi$ s for the KE methods and TCC decreased as test lengths increased from 25 to 100 items for the high average discrimination level. However, at the medium and low average discrimination levels the $RMSE_\pi$ s for the KE and TCC methods decreased only when test lengths increased from 50 to 100 items.

KE (CE) and KE (PSE) performed very similar to each other regardless of the condition. In contrast, TCC did not perform as well as KE (CE) and KE (PSE) at the 25-item test length for both percent anchor and at the 100-item and 30% anchor level. Nevertheless and irrespective of

equating technique, the proportion error was never greater than .131. Specifically, for TCC, the maximum $RMSE$ for the estimated domain expected true scores was 0.131, whereas for KE it was 0.090 (CE) and 0.089 (PSE); all occurred at the 25-item length. Minima were 0.016, 0.012, and 0.014 for TCC, KE (CE), and KE (PSE), respectively. Figure 1 presents some typical $RMSE_{\pi}$ plots for various test lengths, percent anchors, and discrimination levels for a sample size of 1,000; the 2,000 and 10,000 sample size conditions showed virtually the same patterns.

Comparison of $Bias_{\pi}$ as a function of domain score showed that the constant reliability and the nonconstant reliability levels for each cell showed extremely similar performance patterns. Figure 2 contains the $Bias_{\pi}$ plots corresponding to the conditions shown in Figure 1. As can be seen in TCC, $Bias_{\pi}$ decreased as test lengths increased with the greatest bias occurring for the largest portion of the scale at the 25-item length. In addition, the $Bias_{\pi}$ pattern changed from overestimation throughout the continuum to an over- and underestimation pattern as the average discrimination level decreased. This over- and underestimation pattern was also typical of the 50-item length regardless of average discrimination level and percent anchor condition and the 100-item test length for the low average discrimination level. For the 100-item length TCC, $Bias_{\pi}$ showed a slight overestimation bias at the high average discrimination level, and this bias increased when the average discrimination level decreased to medium condition. In contrast, KE (CE) and KE (PSE) showed minimal bias throughout the continuum regardless of average discrimination level, test length, and percent anchor condition.

Summary and Discussion

This study compared TCC, KE (CE), and KE (PSE) using a NEAT design. Generally speaking, KE (CE) and KE (PSE) performed better than or equal to TCC. Differences between KE (CE) and KE (PSE) were very small, and for all intents and purposes, the two approaches could be said to have yielded similar results. Previous work (e.g., Grant, Zhang, & Damiano, 2009) have found similar findings with respect to KE (CE) and KE (PSE).

Using empirical data, Ricker and von Davier (2007) found that external anchor test length affected bias when anchor test length decreased from 20 items to 16 items (approximately 13% of the test length), but that KE (PSE) with optimal bandwidths was less sensitive to this change than the other methods investigated. The current study showed that with tests longer than those used by Ricker and von Davier that external anchor test length had virtually no effect on $RMSE_{\pi}$ and $Bias_{\pi}$ for KE (CE) and KE (PSE). This was also almost always the case with TCC. As would be expected, increasing the test length led to more accurate person estimates. It had been hypothesized that the benefits of longer anchor tests were, in part, due to the increased reliability that one typically sees as test length increases. Support for this hypothesis was not found in the context of external anchor tests.

As mentioned above, the KE results are on the number correct metric, whereas the TCC results are on a logistic ability scale (Haebara, 1980; Stocking & Lord, 1982). Thus, to compare the TCC equating results with those of KE, the equated $\hat{\theta}_{Y|X}$ s (i.e., based on the linked Form Y item parameter estimates) were converted to their corresponding $\hat{\tau}_{TCC}$ s and compared with the parametric expected true scores (i.e., based on the item parameters and θ_Y). This nonlinear transformation of $\hat{\theta}_{Y|X}$ to $\hat{\tau}_{TCC}$ results in some regions of the logistic metric being elongated, whereas other regions are contracted. Given TCC's past performance (e.g., Baker & Al-Karni, 1991) and the fact that TCC equating is essentially a linear transformation, one may ask if the nonlinear transformation of $\hat{\theta}_{Y|X}$ to $\hat{\tau}_{TCC}$ had an impact on the presentation of the TCC results. To address this question, the Form Y response data along with the Form X item parameter estimates were used to estimate theta and compare them with the equated estimated theta $\hat{\theta}_{Y|X}$ using the $RMSD$ and the signed difference statistic (i.e., a comparison of $\hat{\theta}_{Y|X}$ and the estimated

Figure 1. $RMSE_{\pi}$ for $N = 1,000$, nonconstant reliability.
Note. TCC: Circle; τ_{KE-CE} : X; τ_{KE-PSE} : Square. RMSE = root mean square error; TCC = test characteristic curve.

Figure 2. $Bias_{\pi}$ for $N = 1,000$, nonconstant reliability.
Note. TCC: Circle; τ_{KE-CE} : X; τ_{KE-PSE} : Square. TCC = test characteristic curve.

Figure 3. TCC—*RMSD* and *signed difference (SDIFF)* for $N = 1,000$, nonconstant reliability.
Note. TCC = test characteristic curve.

theta if the simulee had taken Form X). Figure 3 Panels A and B show that the *RMSD* between these two estimates is less than 0.0515 logits for the 25-item test length/20% anchor and less than 0.0165 for the 50-item test length/20% anchor; these figures are typical of those seen in the other conditions. The signed difference statistic plots (Panels C and D) show that on the logistic scale, the discrepancies between estimators were small, particularly at the longer test lengths. As such, TCC's ability performance is comparable with past studies when the comparisons are made on the logistic scale. This study's results seem to indicate that although TCC performs well on the theta metric, the nonlinear transformation of $\hat{\theta}$ to the true score metric may exacerbate discrepancies that may lead to incorrect decisions about examinees when based on the true score metric.

Past research has shown that KE compares favorably with linear and equipercentile methods (Mao et al., 2006; von Davier et al., 2006). However, despite the comparability of KE and equipercentile equating one advantage of KE is that it is flexible enough to handle atypical scores in the distributions that equipercentile cannot handle as well (Cid & von Davier, 2015). Similarly, comparisons of KE (CE) or KE (PSE) with TCC have shown that the particular KE method studied performed acceptably vis à vis TCC.

Similar to Godfrey (2007) and Meng (2012), this study used simulated IRT-model-based data that, although it allowed an assessment of equating accuracy, favored the TCC method. As such, it represented a best case scenario for TCC. Thus, in those cases with poor(er) model–data fit, it should be expected that TCC will not perform better (and most likely it will be worse) than it does in our best case scenario (i.e., when there is perfect model–data fit). In fact, previous work (Norman Dvorak, 2009) that utilized a more “neutral” data generation approach than used here found that, generally speaking, KE performed better than TCC.

Overall, the current study showed that both KE and TCC equating produced accurate results, although KE tended to perform better than TCC on the true score scale across most of the conditions. The results seem to indicate that KE should be strongly considered for most equating situations, particularly in light of its flexibility and, in the appropriate context, when IRT assumptions are not met.

Acknowledgment

The authors wish to thank the anonymous reviewers for their helpful comments that enhanced the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

- Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, 55, 1-25. Retrieved from <http://www.jstatsoft.org/v55/i06/>
- Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.
- Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement*, 15, 78.
- Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating under the generalized kernel equating framework. *Journal of Educational Measurement*, 49, 269-284.
- Cid, J. A., & von Davier, A. A. (2015). Examining potential boundary bias effects in Kernel smoothing on equating: An introduction for the adaptive and Epanechnikov kernels. *Applied Psychological Measurement*, 39, 208-222.

- 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.
- Ferrando, P. J. (2004). Kernel-smoothing estimation of item characteristic functions for continuous personality items: An empirical comparison with the linear and the continuous-response models. *Applied Psychological Measurement, 28*, 95-109.
- Godfrey, K. E. (2007). *A comparison of kernel equating and IRT true score equating methods* (Unpublished doctoral dissertation). University of North Carolina, Greensboro.
- Grant, M. C., Zhang, L., & Damiano, M. (2009). *An evaluation of kernel equating: Parallel equating with classical methods in the SAT Subject Tests[TM] Program* (Technical Report No. 09-06). Princeton, NJ: Educational Testing Service.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares methods. *Japanese Psychological Research, 22*, 144-149.
- Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement, 51*, 201-211.
- Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.
- Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Technical Report No. 89-84). Princeton, NJ: Educational Testing Service.
- Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*, 39-52.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.
- Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*, 121-134.
- Meng, Y. (2012). *Comparison of kernel equating and item response theory equating methods* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.
- Norman Dvorak, R. L. (2009). *A comparison of kernel equating to the test characteristic curve method* (Unpublished doctoral dissertation). University of Nebraska-Lincoln.
- Ricker, K. L., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design* (ETS RR-07-44). Princeton, NJ: Educational Testing Service.
- Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*, 249-275.
- Stocking, M. L., & Lord, F. M. (1982). *Developing a common metric in item response theory* (ETS RR-82-25). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.
- 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 pseudotests constructed from real test data* (ETS RR-06-02). Princeton, NJ: Educational Testing Service.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and poststratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15-32.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. New York, NY: Springer-Verlag.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG for Windows: Multiple-group IRT analysis and test maintenance for binary items (Version 3. 0) [Computer software]. Chicago, IL: Scientific Software International.