Article

How Important is the Choice of Bandwidth in Kernel Equating?

Applied Psychological Measurement
2021, Vol. 45(7-8) 518–535

DOI: 10.1177/01466216211040486
journals.sagepub.com/home/apm

Gabriel Wallin[1], **Jenny Häggström**[1], and **Marie Wiberg**[1]

Abstract

Kernel equating uses kernel smoothing techniques to continuize the discrete score distributions when equating test scores from an assessment test. The degree of smoothness of the continuous approximations is determined by the bandwidth. Four bandwidth selection methods are currently available for kernel equating, but no thorough comparison has been made between these methods. The overall aim is to compare these four methods together with two additional methods based on cross-validation in a simulation study. Both equivalent and non-equivalent group designs are used and the number of test takers, test length, and score distributions are all varied. The results show that sample size and test length are important factors for equating accuracy and precision. However, all bandwidth selection methods perform similarly with regards to the mean squared error and the differences in terms of equated scores are small, suggesting that the choice of bandwidth is not critical. The different bandwidth selection methods are also illustrated using real testing data from a college admissions test. Practical implications of the results from the simulation study and the empirical study are discussed.

Keywords

kernel equating, continuization, bandwidth selection, evaluation

Introduction

Kernel equating (KE) is an observed-score equating framework aiming at making test scores from standardized tests comparable between administrations (von Davier et al., 2004). Based on the scores from two test administrations, the objective is to find equivalent scores in terms of the latent trait the test is constructed to measure. Following the Braun and Holland (1982) definition of equivalent scores, KE makes use of the equipercentile transformation to equate the test scores. It is functionally composed of two cumulative distribution functions (CDFs), each representing the respective distribution of the test scores to be equated. For the equipercentile transformation to be properly defined, these functions need to be continuous and monotonically increasing. This is

[1]Department of Statistics, USBE, Umeå University.

Corresponding Author:
Gabriel Wallin, Department of Statistics, Umeå University, Biblioteksgränd 6, 901 87 Umeå 90187, Sweden.
Email: gabriel.wallin@umu.se

generally not true since test scores most often are discrete. For this reason, KE employs smoothing techniques where a, usually Gaussian, kernel function approximates the discrete CDFs with continuous functions. Regardless of the choice of kernel function (e.g. Gaussian, uniform, and logistic) one needs to select a bandwidth which determines the smoothness of the continuous approximations and ultimately, the equated scores. Since undersmoothing results in estimated distributions that suffer from excessive sampling noise and oversmoothed distributions will blur the characteristics of the underlying density, it is of interest to investigate to what extent the choice of bandwidth influences the KE estimator.

The overall aim of this study is to examine if and how the bandwidth choice affects the equated scores, and if specific bandwidth selection methods are more suitable for certain test scenarios. The bandwidth selection in KE is particularly interesting to investigate since it defines the main difference to traditional equating methods. If the influence of the bandwidth on the equated scores is strong, it is important to know which bandwidth selection method to use. If the choice is not sensitive, it could encourage practitioners that are lacking strong theoretical training to consider using KE. In the KE literature to date, four different bandwidth selection methods have been proposed: the penalty method (von Davier et al., 2004), the double smoothing (DS) method (Häggström & Wiberg, 2014), the cross-validation (CV) method (Liang & von Davier, 2014), and the Silverman's rule of thumb (SRT) method (Andersson & von Davier, 2014). For clarity, the cross-validation method in Liang and von Davier (2014) is hereinafter referred to as the likelihood cross-validation method (LiCV). In Häggström and Wiberg (2014), comparisons between the DS and penalty method showed slight differences in terms of mean squared error (MSE) of the estimated mean of the equated scores. The largest differences were seen for skewed score distributions. They considered symmetric and skewed data, using both the equivalent groups (EG) design and the non-equivalent groups with anchor test (NEAT) design. In Liang and von Davier (2014), comparisons between the LiCV method and the penalty method showed small differences in terms of bias of the density estimate, but the former was the preferred choice for symmetric data. They considered the EG design for both symmetric, skewed, and bimodal data. In Andersson and von Davier (2014), comparisons between the SRT and penalty method showed great similarities in terms of the equated scores. They considered symmetric and skewed data under both the EG and NEAT design. The DS, LiCV, and SRT methods have thus only been compared with the penalty method and never with each other. Furthermore, the previous studies on bandwidth selection in KE have used different data collection designs and evaluation criteria, making it even harder to compare the results between the studies.

Outside the KE literature, leave-one-out cross-validation (LCV) has been widely discussed in density estimation, see for example Jones et al. (1996), Sheather (2004), and Wasserman (2006). LCV is often used as a benchmark method for novel bandwidth selection methods within the kernel density and kernel regression frameworks, see for example Park and Marron (1990) and Häggström and De Luna (2010). Thus, in addition to the four currently available bandwidth selection methods in KE, LCV as well as a penalized LCV are included in the comparison for completeness.

The six bandwidth methods will be evaluated and compared with each other in a simulation study where the test length, number of test takers, and distributions of the test scores are varied for both the EG and NEAT design. All methods will also be illustrated empirically with real test data from a college admissions test.

The rest of the paper is structured as follows. First, a brief review of KE is given, then the six bandwidth selection methods are described. This is followed by the simulation study and the empirical illustration. The paper is concluded with a discussion together with some practical recommendations.

The Kernel Equating Framework

KE comprises five steps: (1) Presmoothing the score distributions; (2) Estimating the score probabilities; (3) Continuizing the estimated score distributions; (4) Equating; and (5) Evaluating the estimated equating function (e.g., by calculating the standard error of equating [SEE]; von Davier et al. (2004); González and Wiberg (2017)). In this paper, the third step, for which KE offers a unique solution in comparison with other equating methods, will be examined.

The test scores from test forms X and Y are denoted by X and Y, respectively, with realizations $x_j, j = 1, \ldots, J$ and $y_k, k = 1, \ldots, K$. The scores X and Y are viewed as random variables with CDFs $F_X(\cdot)$ and $G_Y(\cdot)$, respectively. In the NEAT design, anchor scores A with realizations $a_l, l = 1, \ldots, L$, are also measured. The equipercentile transformation $\varphi_Y(x)$ that equates test form X to test form Y is defined as

$$y = \varphi_Y(x) = G_Y^{-1}(F_X(x)) \tag{1}$$

To define the KE estimator of $\varphi_Y(x)$ in equation (1), we introduce the following notation: Let σ_X^2 denote the variance of X, $r_j = \Pr(X = x_j|T)$, the x_j score probability on the target population T, and Z a standard normal random variable. Define $a_X^2 = \sigma_X^2/(\sigma_X^2 + h_X^2)$, where h_X is the bandwidth of the X score kernel density estimate and $\mu_X = \sum_j x_j r_j$. Using a Gaussian kernel, the discrete score variable X is replaced by $X(h_X) = a_X(X + h_X Z) + (1 - a_X)\mu_X$, where $X(h_X)$ is a continuous random variable constructed to preserve the first two moments of X. By letting $\mathbf{r} = (r_1, \ldots, r_J)^\top$ and $\Phi(z)$ denote the standard normal distribution function, it can be shown that the CDF of $X(h_X)$ is given by

$$\begin{aligned} F_{h_X}(x; \mathbf{r}) &= \Pr(X(h_X) \le x) \\ &= \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right). \end{aligned} \tag{2}$$

The continuized CDF of Y, denoted $G_{h_Y}(y; \mathbf{s})$, is defined analogously using $\mathbf{s} = (s_1, \ldots, s_K)^\top$ and $s_k = \Pr(Y = y_k|T)$. By letting $F_{h_X}(x; \widehat{\mathbf{r}}) = \widehat{F}_{h_X}(x)$ and $G_{h_Y}(y; \widehat{\mathbf{s}}) = \widehat{G}_{h_Y}(y)$, the KE estimator of the equipercentile transformation in equation (1) is defined as

$$\widehat{\varphi}_Y(x) = \widehat{G}_{h_Y}^{-1}\left(\widehat{F}_{h_X}(x)\right) \tag{3}$$

Equations (2) and (3) show the dependence of the equated scores on the bandwidth through their dependence on the continuized score CDFs. Optimal choices of the bandwidths h_X and h_Y would thus find the members of the family of continuous distributions $\{\widehat{F}_{h_X}, h_X > 0\}$ and $\{\widehat{G}_{h_Y}, h_Y > 0\}$ that, composed as $\widehat{\varphi}_Y(x)$, yield the best estimator of $\varphi_Y(x)$.

The most common evaluation measure of the equating estimator given in equation (3) is the SEE (von Davier et al., 2004). The estimated SEE consists of three components; the Jacobian of the estimated equating transformation, denoted $\widehat{J}_{\varphi_Y}$, the Jacobian of the design function which maps the (presmoothed) score distributions into $\mathbf{r}$ and $\mathbf{s}$, denoted $\widehat{J}_{\mathrm{DF}}$, and a matrix $\mathbf{C}$ that relates to the covariance matrix of the (presmoothed) score distributions. The SEE of $\widehat{\varphi}_Y(x)$ is formed by combining these three components and calculating the length of the resulting vector, that is,

$$\mathrm{SEE}_Y(x) = \left\| \widehat{\mathbf{J}}_{\varphi_Y} \widehat{\mathbf{J}}_{DF} \mathbf{C} \right\| \tag{4}$$

Another common measure is the Percent Relative Error (PRE; von Davier et al. 2004), which measures the discrepancy between the p:th moment of the equated scores and that of the Y scores. Letting $\mu_p(Y) = \sum_k (y_k)^p s_k$ and $\widehat{\mu}_p(\varphi_Y(X)) = \sum_j (\widehat{\varphi}_Y(x_j))^p r_j$, the PRE is defined as

$$\mathrm{PRE}(p) = 100\left(\frac{\widehat{\mu}_p(\varphi_Y(X)) - \mu_p(Y)}{\mu_p(Y)}\right)$$

Bandwidth Selection Methods in Kernel Equating

We consider data-driven selection of one bandwidth per test score density estimator. Note, it is also possible to manually select bandwidths that fulfill certain objectives. For example, when selecting a very large bandwidth the KE estimator is similar to the linear equating transformation, and when setting the bandwidths equal to 0.33 the KE estimator will approximate the traditional equipercentile transformation that uses linear interpolation (von Davier et al., 2004). Another possibility is to use adaptive kernels (González & von Davier, 2017) which allow for different bandwidths along the data points. All expressions in this section are in terms of the X scores, but expressions for the Y scores are analogous.

The Penalty Method

The most common way of selecting the bandwidth in KE is by minimizing the sum of the squared distances between the estimated score probabilities $\widehat{r}_j$, $j = 1, \ldots, J$, and the estimated density function $\widehat{f}_{h_X}(x_j)$, $j = 1, \ldots, J$, where $\widehat{f}_{h_X}$ denotes the derivative of $\widehat{F}_{h_X}$. To ensure smoothness, a penalty function can be added to the loss function which prevents the estimated density from exhibiting large fluctuations. The penalty method thus selects the bandwidth that minimizes

$$\mathrm{PEN}(h_X) = \sum_j \left(\widehat{r}_j - \widehat{f}_{h_X}(x_j)\right)^2 + \kappa \cdot \sum_j A_j \tag{5}$$

where $A_j = 1$ if

$$\left[\left(\widehat{f}'_{h_X}(x_j - w) > 0\right) \cap \left(\widehat{f}'_{h_X}(x_j + w) < 0\right)\right]$$

or

$$\left[\left(\widehat{f}'_{h_X}(x_j - w) < 0\right) \cap \left(\widehat{f}'_{h_X}(x_j + w) > 0\right)\right],$$

and $A_j = 0$ otherwise (Lee & von Davier, 2011; von Davier, 2013). The term κ is a weight that determines the size of each penalty, $\widehat{f}'_{h_X}(x_j)$ is the derivative of $\widehat{f}_{h_X}(x_j)$, and w is a constant that determines the neighborhood of x_j for which the penalty function will penalize choices of h_X that let $\widehat{f}'_{h_X}(x_j)$ change sign. Typically $w = 0.25$ (see e.g., von Davier et al. (2004), Häggström and Wiberg (2014) and Andersson and von Davier (2014)).

Silverman's Rule of Thumb

A common loss function when selecting bandwidth in density estimation is the asymptotic mean integrated squared error (AMISE; Jones et al., 1996). For a normally distributed random variable, minimizing the AMISE with respect to the bandwidth results in the approximation known as Silverman's rule of thumb (Scott, 1992). Andersson and von Davier (2014) implemented this bandwidth for KE which, adjusted for a_X, equals

$$\mathrm{SRT}(h_X) = \frac{9\sigma_X}{\sqrt{100n_X^{2/5} - 81}}.$$

Double Smoothing

DS was introduced by Hall et al. (1992) for nonparametric density estimation and implemented within KE by Häggström and Wiberg (2014). Within KE, the procedure starts by using a large, subjectively chosen pilot bandwidth q_X to estimate f_{q_X} at the score values and the values halfway between them, that is, at the points $\mathbf{x}^* = \{x_l^*\} = [x_1, x_1 + 0.5, x_2, \ldots, x_J - 0.5, x_J]^\top$, $l = 1, \ldots, 2J - 1$. Next, f_{h_X} is estimated at $\mathbf{x}^*$ using $\widehat{f}_{q_X}$ at the actual score values $\mathbf{x} = [x_1, x_2, \ldots, x_J]^\top$ instead of using the estimated score probabilities $\widehat{r}_j$. Thus, a DS estimate $\widehat{f}^*_{h_X}$ is obtained. The bandwidth that minimizes the sum of the squared difference between the lth DS estimate $\widehat{f}^*_{h_X}(x)$ and $\widehat{r}^*_l$ is selected, where

$$\widehat{f}^*_{h_X}(x) = \sum_{j=1}^{J} \widehat{f}_{q_X}(x_j)\phi\left(\frac{x - \widehat{a}_X x_j - \left(1 - \widehat{a}_X\right)\widehat{\mu}_X}{h_X \widehat{a}_X}\right) \frac{1}{h_X \widehat{a}_X},$$

$\phi(z)$ denotes the standard normal density function,

$$\widehat{f}_{q_X}(x) = \sum_{j=1}^{J} r_j \phi\left(\frac{x - \widehat{a}_X^{q_X} x_j - \left(1 - \widehat{a}_X^{q_X}\right)\widehat{\mu}_X}{q_X \widehat{a}_X^{q_X}}\right) \frac{1}{q_X \widehat{a}_X^{q_X}},$$

with $\widehat{a}_X^{q_X} = \sqrt{\widehat{\sigma}_X^2 / (\widehat{\sigma}_X^2 + q_X^2)}$ and $\widehat{r}^*_l = \widehat{r}^*_{l+1/2}$ if l is even and $\widehat{r}^*_l = \widehat{f}^*_{h_X}(x^*_l)$ if l is odd.

The DS criterion can be written as

$$\mathrm{DS}(h_X) = \sum_{l=1}^{2J-1} \left(\widehat{r}^*_l - \widehat{f}^*_{h_X}(x^*_l)\right)^2.$$

The Likelihood Cross-Validation Method

LiCV applied to KE was suggested by Liang and von Davier (2014), and their method of bandwidth selection starts by randomly splitting the data into two subsamples. The first subsample is used to estimate a set of Gaussian kernel densities,

$$\widehat{f}_{h_X}^{(1)} = \sum_j \widehat{r}_j \phi\left(\frac{x - \widehat{a}_X x_j - \left(1 - \widehat{a}_X\right)\widehat{\mu}_X}{\widehat{a}_X h_X}\right)$$

for a set of bandwidths $h = [0.01, 0.02, \ldots, 5]$, where the "(1)" notation indicates that the quantities are calculated using only the first subsample. The density for each value of h is then used as an intensity parameter in a set of Poisson likelihood functions, where the score frequencies are taken from the second subsample. The value of h that maximizes the likelihood function is stored. The criterion of the LiCV method can be expressed as

$$\text{LiCV}(h_X) = \max_h L\left(n_{x_j}; \widehat{f}_{h_X}^{(1)}\right) = \max_h \prod_{j=1}^{J} \frac{e^{-N_X^{(1)}\widehat{f}_{h_X}^{(1)}(x_j)}\left(N_X^{(1)}\widehat{f}_{h_X}^{(1)}(x_j)\right)^{n_{x_j}^{(2)}}}{n_{x_j}^{(2)}!}$$

where $N_X^{(1)}$ is the number of test takers in the first subsample, and $n_{x_j}^{(2)}$ is the number of test takers with $X = x_j$ in the second subsample. This procedure of randomly splitting the data set and selecting the bandwidth that maximizes the Poisson likelihood function is repeated 1000 times and the median of the resulting 1000 bandwidths is selected as the optimal bandwidth.

Penalized Leave-One-Out Cross-Validation

There are two objectives when estimating the distribution of X (h_X); $\widehat{f}_{h_X}$ should both be a good estimate of the true density f but also track the shape of the relative score frequencies. Regarding the first objective, Stone (1984) showed that

$$\frac{\int \left(f(x) - \widehat{f}_{h_{CV}}(x)\right)^2 dx}{\inf_h \int \left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{a.s.} 1$$

where $\widehat{f}_{h_{CV}}(x)$ is the kernel density estimator of f with bandwidth h_{CV} selected using LCV. To make sure that the estimated density of X (h_X) also tracks the estimated probabilities the following criterion, for a Gaussian kernel, can be minimized

$$\text{LCV}(h_X) = \frac{1}{J}\sum_{j=1}^{J}\left(\widehat{r}_j - \widehat{f}_{h_X}^{-j}(x_j)\right)^2 \tag{6}$$

where

$$\widehat{f}_{h_X}^{-j}(x_j) = \sum_{\substack{l=1 \\ l \neq j}}^{J} \widehat{r}_l \phi\left(\frac{x_j - \widehat{a}_X x_l - \left(1 - \widehat{a}_X\right)\widehat{\mu}_X}{h_X \widehat{a}_X}\right) \frac{1}{h_X \widehat{a}_X},$$

is the estimate of $f(x_j)$ based on the subsample with $(x_j, \widehat{r}_j)$ left out. The estimated quantities $\widehat{a}_X$ and $\widehat{\mu}_X$ are based on the full sample. Note that the expression in equation (6) is analogous to the first term in equation (5), the criterion of the penalty method. By the same argument used to motivate the second term of the penalty method, we propose to modify the LCV method by adding the penalty function A_j. A penalized LCV criterion is thus defined as

$$\mathrm{PLCV}(h_X) = \frac{1}{J}\sum_{j=1}^{J}\left(\hat{r}_j - \hat{f}_{h_X}^{-j}(x_j)\right)^2 + \kappa \cdot \sum_{j=1}^{J} A_j.$$

Simulation Study

A simulation study is conducted under both the EG and NEAT design to evaluate 1) how big the differences are between the bandwidths described in the previous section and 2) if any such differences are reflected in the equated scores. Most of the presented results are based on the NEAT design since it is a very common design in practice. Additionally, the EG results are often in line with those of the NEAT design except when indicated.

Simulation Design

All simulations are repeated with 1000 iterations each, with sample sizes of $n = \{100, 1000, 5000\}$, test lengths of $J - 1 = K - 1 = \{40, 80\}$, and anchor test lengths of $L - 1 = \{20, 40\}$. For the two smallest sample sizes, both test lengths $\{40, 80\}$ are considered in combination with both anchor test lengths $\{20, 40\}$, and for $n = 5000$, a test length of 80 together with an anchor test length of 40 is considered. By altering the test lengths in this fashion, it is possible to explore how the equating function is affected by a changing number of observed-score frequencies, both on the main test and on the anchor test. It should be noted that relatively long tests can suffer from other issues as well, like a changing shape of the score distributions and a weaker correlation between the anchor and the test scores. These factors are assumed to be negligible in this study.

Data generation and all computations are performed with the software R (R Core Team, 2018) and the R package **kequate** (Andersson et al., 2013) is used for kernel equating. R code for the simulation study can be obtained from the corresponding author upon request.

The data generating process (DGP) described below was chosen in an attempt to mimic the characteristics of real testing data. The scores of the test takers from population P who are given test form X are denoted X and the scores of the test takers from population Q who are given test form Y are denoted Y, where we consider number-correct scoring. In the EG design of this study, the two samples of test takers are only randomly different from each other, that is, $P = Q$. In the NEAT design, $P \neq Q$ and a population weight of 0.5 is used for the target population, that is, $T = 0.5P + 0.5Q$. Since previous studies on bandwidth selection in KE have used post-stratification in the NEAT design to form the equipercentile transformation function (von Davier et al., 2004), this is the approach here as well to allow the results to be more easily compared. We now describe the DGP for test form X.

1. Generating true score probabilities $r_j = \Pr(X = x_j)$ and $p_{jl} = \Pr(X = x_j \cap A = a_l)$, $j = 1, \ldots, J$, $l = 1, \ldots, L$.

Generate auxiliary variable(s) according to:
For EG

$$U_i \sim \mathrm{Beta}(\alpha, \beta),$$

And for NEAT

$$U_i, V_i \sim \text{Normal copula bivariate distribution with Beta}(\alpha, \beta) \text{ marginal, correlation set to } \rho = 0.75 \text{ and } i = 1,\ldots,n.$$

Then individual scores, to be used for generating r_j and p_{jl}, are calculated by rounding the auxiliary variable(s) times the test length to the nearest integer, $X_i^* = \lfloor (J-1)U_i \rceil$ and for NEAT also $A_i^* = \lfloor (L-1)V_i \rceil, \quad i = 1,\ldots,n$. We use a combination of floor and ceiling notation to denote rounding to the nearest integer. The floor function of a variable x, $\lfloor x \rfloor$, returns the greatest integer less than or equal to x, and the ceiling function $\lceil x \rceil$ returns the smallest integer greater than or equal to x. Under both the EG and NEAT design, the shape parameters for the beta distributions are set to $(\alpha, \beta) = \{(5, 5), (5, 2), (2, 5)\}$ to produce symmetric, negatively skewed and positively skewed score data, respectively. To produce bimodal score data, a mixture of Beta distributions with $(\alpha, \beta) = (25, 15)$ and $(\alpha, \beta) = (15, 25)$ is used. In the NEAT design, the correlation between the test score and anchor test score is set to $\rho = 0.75$, since it operationally has been standard practice to aim for anchor tests with strong correlation to the total test scores. The R package **copula** (Yan, 2007) is used to generate data from a Normal copula bivariate distribution with Beta marginals.

Let $\mathbf{n}_J^* = \{n_j^* = \sum_{i=1}^n I(X_i^* = x_j)\}$ and $\mathbf{n}_{JL}^* = \{n_{jl}^* = \sum_{i=1}^n I(X_i^* = x_j \cap A_i^* = a_l)\}$, $j = 1, \ldots, J$ and $l = 1, \ldots, L$. For the EG design, log-linear models regressing n_j^* on a function of x_j are fitted. The Akaike information criterion (AIC; Akaike, 1998) is used for model selection under the EG design and the Bayesian information criterion (BIC; Schwarz, 1978) is used for model selection under the NEAT design, aiming at parsimonious models. The choice of criteria are based on previous research where the AIC has been shown to be a suitable model-fit measure for univariate log-linear models (Moses & Holland, 2009) and the BIC has been shown to be effective for bivariate smoothing (Moses & Holland, 2010). The fitted values from the AIC/BIC selected models are used as true score probabilities. A similar procedure is used to generate the fitted probabilities under the NEAT design (for details, see Section 11.1 in von Davier et al., 2004). The procedure of using fitted values from estimated log-linear models as true score probabilities follows the approach in both Liang and von Davier (2014) and Häggström and Wiberg (2014). Note that this step is only performed once (per simulation configuration) and the same true probabilities are then used in all simulation iterations.

2. Generating test score frequencies

For each simulation iteration, using the true score probabilities generated in step 1, we generate test score sample frequencies as

$$\mathbf{n}_J \sim \text{Multinomial}(n, (r_1,\ldots,r_J)^\top) \text{ and } \mathbf{n}_{JL} \sim \text{Multinomial}(n, (p_{11},\ldots,p_{JL})^\top)$$

The DGP for test form Y is analogous to that of test form X with exception that under the NEAT design, the data from population Q is shifted by five units along the score axis. This means that for the symmetric, negative, and bimodal distributional scenarios the test form taken by the Q sample is more difficult than that taken by the P sample, and vice versa for the positive distributional scenario. Figure 1 illustrates the score distributions considered under the EG design.

For the EG design, the log-linear models in step 1 in the DGP preserved the first two and three moments, respectively, of the X and Y scores for the symmetric and skewed distributions, and the first three moments for the bimodal distributions. For the NEAT design, the models preserved the first four moments of the X, Y, and A scores, respectively, and the first and second cross-moment for the symmetric and negatively skewed distributions. The models for the positively skewed and bimodal distributions preserved the first two moments of the X, Y, and A scores, and the first cross-moment. An alternative would have been to preserve the same number of moments for the

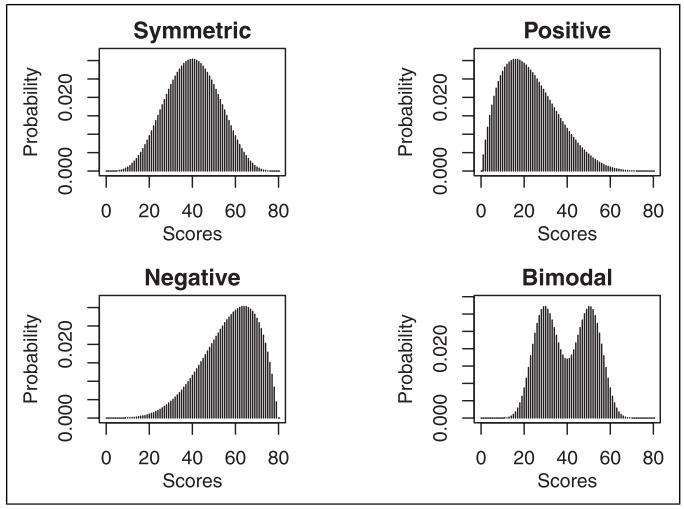

Figure 1. The distributional settings under the EG design in the simulation study.

different data collection designs. However, we have chosen to use the best fitting model according to the AIC/BIC which we believe to be a better reflection of equating done in practice. Lastly, since it is very common to presmooth the score distributions, the sampled data from step 2 were presmoothed in each simulation iteration using the same models as the ones used to generate the true score probabilities. Finally, note that since the populations of test takers have been created using two random samples from the beta distribution, the identity function is not the true equating function for any of the considered data collection designs.

Evaluation Criteria

To evaluate the equating results generated by the different bandwidth methods, a comparison between the distribution of $\widehat{\varphi}_Y(X)$, the estimated KE transformation evaluated at the discrete X score points, and the distribution of Y was made for each KE estimator. Let $\mu_Y = \sum_k y_k s_k$, $\widehat{\mu}_Y = \sum_j \widehat{\varphi}_Y(x_j) r_j$ and let $\widehat{\mu}_Y^{(g)} = \sum_j \widehat{\varphi}_Y^{(g)}(x_j) r_j$ denote the estimator of the mean based on the gth replicate. The MSE of $\widehat{\mu}_Y$ over 1000 replications was calculated as

$$\mathrm{MSE}\left(\widehat{\mu}_Y\right) = \left[\frac{1}{1000}\sum_{g=1}^{1000}\left(\widehat{\mu}_Y^{(g)} - \mu_Y\right)\right]^2 + \frac{1}{1000-1}\sum_{g=1}^{1000}\left[\widehat{\mu}_Y^{(g)} - \frac{1}{1000}\sum_{g=1}^{1000}\widehat{\mu}_Y^{(g)}\right]^2$$

Letting $\overline{\varphi}_Y(x_j) = \frac{1}{1000}\sum_{g=1}^{1000}\widehat{\varphi}_Y^{(g)}(x_j)$, the standard error (SE) of $\widehat{\varphi}_Y(x_j)$ was calculated as

$$\mathrm{SE}\left(\widehat{\varphi}_Y(x_j)\right) = \sqrt{\frac{1}{1000-1}\sum_{g=1}^{1000}\left(\widehat{\varphi}_Y^{(g)}(x_j) - \overline{\varphi}_Y(x_j)\right)^2}$$

For both the MSE and SE, the corrected sample standard deviation formula is applied, for which the squared distances are divided by $G - 1 = 1000 - 1$. For each method and scenario, the PRE of the first 10 moments were also calculated. Furthermore, the mean equating transformation of the 1000 replicates was calculated for every estimator together with the difference that matters

(DTM; Dorans & Feigenbaum, 1994) which is referring to all differences larger than half a raw score unit.

Lastly, to judge the validity of the analytical SEEs in equation (4), bootstrap standard errors were calculated under the EG design with symmetric data in an additional simulation. Here, the sample size was n = 10,000, the test length was 40 and no presmoothing was conducted. 1000 bootstrap samples were drawn from each data set.

Simulation Results

The mean of the bandwidths for each scenario and method were calculated and are found in the supplemental material. Generally speaking, for a given scenario, the bandwidth methods result in very different bandwidths. For example, for the symmetric scenario under the EG design, using 80 items and a sample size of 100, the smallest mean bandwidth for the X scores is 0.34 (LCV) and the largest mean bandwidth equals 4.84 (SRT). For the negatively skewed data under the NEAT design with 1000 test takers, 80 items and 20 anchor items, the mean bandwidths for the X scores were 0.59 (Penalty), 2.96 (SRT), 0.57 (DS), 3.09 (LiCV), 0.32 (LCV), and 1.05 (PLCV). This kind of spread between the different bandwidth selection methods were typical for all scenarios, see the supplemental material for the full table. The variances of the bandwidths were also calculated and the differences between the methods were mostly small. Generally, the variances of the SRT and PLCV methods were the largest under the EG design and the variance of the LiCV method were the largest under the NEAT design. The LCV method had the lowest variance for every scenario and data design.

Table 1 shows the MSE of $\widehat{\mu}_Y$ for the symmetrical distribution scenario under the NEAT design. For a given sample size the MSE increases when the test length increases, and for a given test length, the MSE decreases when the sample size increases. For these sample size and test length effects, the MSE is decreasing by more when the sample size grows from 100 to 1000 than what it is increasing when the number of items grows from 40 to 80. Furthermore, when the anchor test length increases from 20 to 40 items, the MSE decreases for all sample sizes and both data collection designs. For a sample size of 1000, the MSE is more than halved when the anchor test length is doubled. Notably, when the number of test takers equals 5000, the test length equals 80, and the number of anchor items equal 40, the MSE is down to the same size as for the scenario with 1000 test takers, 40 items, and 20 anchor items. For the other distributional scenarios, the results are in line with those of the symmetric distributional scenario. However, the MSE is generally smaller regardless of distribution under the NEAT design compared to the EG design. The MSE is smallest when the data are bimodal, with the overall smallest MSE under the NEAT design with

Table 1. The MSEs for all symmetric distributional scenarios considered under the NEAT design. The asterisk (*) indicates that the number of anchor items equals 40, otherwise they equal 20.

Distribution	Design	Penalty	SRT	DS	LiCV	LCV	PLCV
Symmetric	n = 100, $J - 1$ = 40	0.329	0.326	0.328	0.326	0.329	0.327
	n = 100, $J - 1$ = 80	1.638	1.624	1.638	1.627	1.638	1.637
	n = 1000, $J - 1$ = 40	0.032	0.032	0.032	0.032	0.029	0.032
	n = 1000, $J - 1$ = 80	0.111	0.111	0.111	0.111	0.111	0.111
	n = 100, $J - 1$ = 80*	1.015	1.024	1.023	1.025	1.021	1.029
	n = 1000, $J - 1$ = 80*	0.046	0.046	0.046	0.046	0.046	0.046
	n = 5000, $J - 1$ = 80*	0.031	0.031	0.031	0.031	0.030	0.031

bimodal data, a sample of 1000 test takers and 40 items. The full table can be viewed in the supplemental material.

Figure 2 shows the performance of each KE estimator in terms of PRE. The results are displayed for the NEAT design using a sample size of 1000 test takers, a test length of 80, and an anchor test length of 20. The PRE for all other sample sizes, test lengths and data collection designs are available and can be found in the supplemental material. For the symmetric distribution, the PREs are very similar and by far the largest in magnitude. For the negatively skewed distributions, the LiCV method is clearly outperformed by the other methods with the SRT method being second worst. In contrast, with positively skewed distributions the LiCV and SRT instead achieve the best results. For the bimodal setting, the differences between the methods are small but with the SRT method performing slightly worse in the mid-range of the scores. The difference in PRE for the other scenarios are generally small. However, the KE estimator using the SRT method is among the worst under the EG design regardless of sample size, test length, and distributional scenario. It is also among the worst for all test lengths and distributional scenarios under the NEAT design when the sample size is small ($n = 100$). It is possible that the relative weak performance of the SRT method is due to its underlying normality assumption which is unrealistic for most test data, including those generated in this simulation study.

In Figure 3, the PRE for the symmetric scenario under the NEAT design is presented when the number of test takers is 5000 per test group, the tests consist of 80 items, and the anchor test length equals 40. As in Figure 2, the differences between the KE estimators are small. In the four last moments, there is a visible, although small, difference between the LiCV-based estimator and the other estimators. It is also notable that the absolute magnitudes of the PREs are considerably smaller compared to when the number of test takers is 1000 and the anchor length is 20. It suggests that increasing the sample size and the length of the anchor test yields an equating estimator that better approximates the Y score distribution.

Figure 4 displays the simulation SEs for the same scenarios as those presented in Figure 2. The estimators are similar in performance but the LiCV method is the worst for the lowest scores in the symmetric setting and the PLCV method the worst in the tails in the bimodal setting. For the negatively skewed data, there are only small differences between each method; however, for

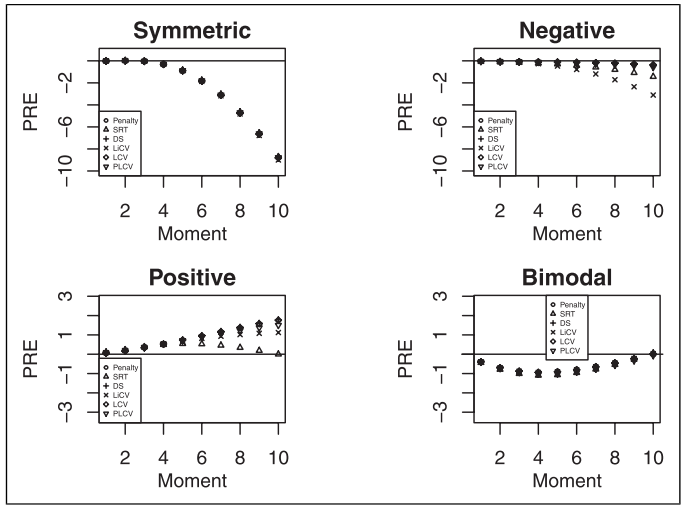

Figure 2. The PRE for every KE estimator under the NEAT design with a sample size of 1000, a test length of 80 and an anchor test length of 20.

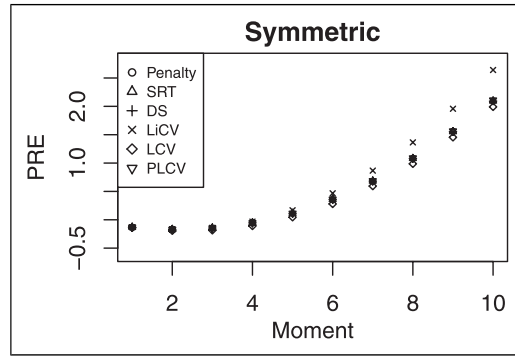

Figure 3. The PRE for every KE estimator for symmetric test score distributions under the NEAT design with a sample size of 5000, a test length of 80, and an anchor length of 40.

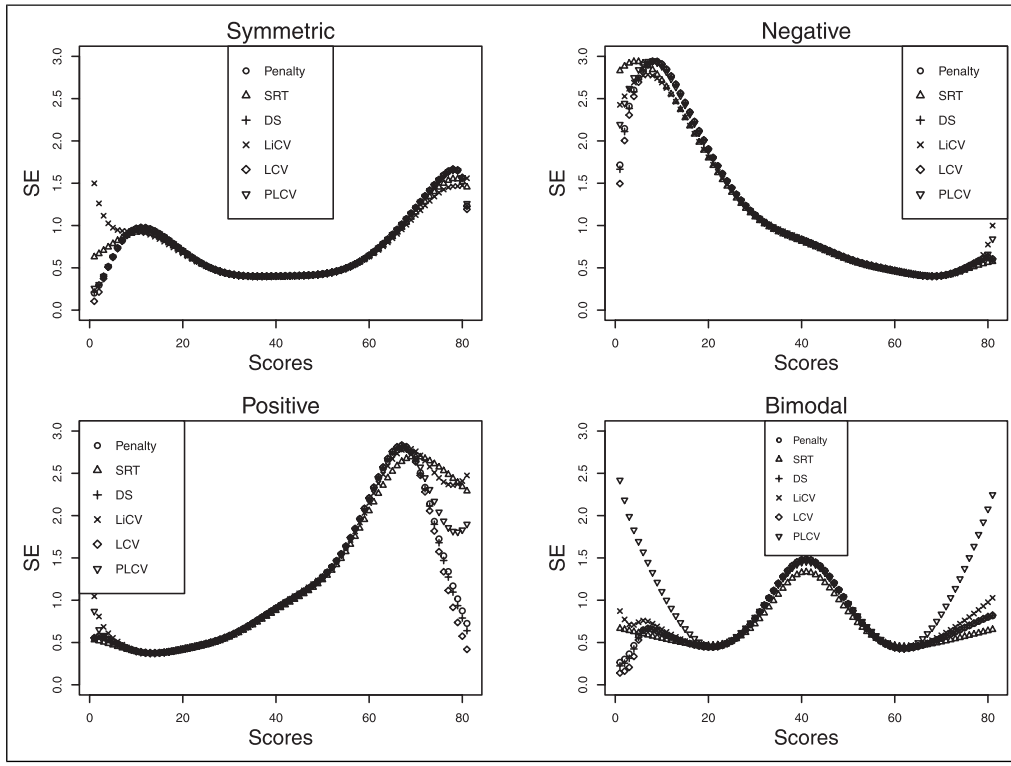

Figure 4. The SE for every KE estimator under the NEAT design with a sample size of 1000, a test length of 80, and an anchor test length of 20.

positively skewed data the LCV method is superior in the top scores. The distributional scenarios are also reflected in the SE; for the negative skew the SE is substantially higher at the lowest scores, and vice versa for the positive skew.

Figure 5 illustrates the SE for all KE estimators in a similar way as in Figure 4, but with 5000 test takers per group and an anchor test length of 40. As in Figure 4, there are mostly small differences between the KE estimators. The exception is the LiCV-based KE estimator which

demonstrates larger SEs in the tails of the score scale, especially for the lowest scores. However, in a practical sense this is often not critical since most sensitive decisions are made at the other end of the score scale. It is interesting to note that there is no apparent difference in the SEs compared to when the sample size is 1000 and the anchor test length is 20, as was evident when comparing the PRE.

For the estimation of the SE, Figure 6 displays the analytical and bootstrap SE for every KE estimator. There are only slight differences between the analytical and bootstrap SEs for most scores, and differences are seen only at the tails. As expected, the SRT-based KE estimator best manages the tails since it accounts for bandwidth variability in the estimation. However, the

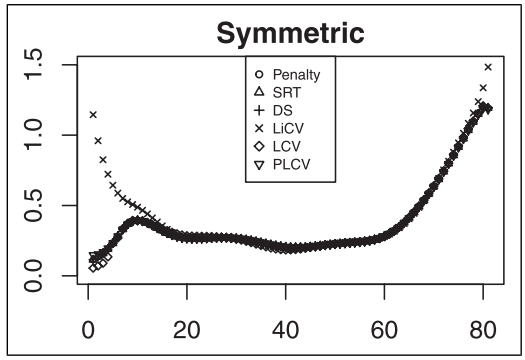

Figure 5. The SE for every KE estimator under the NEAT design with a sample size of 5000, a test length of 80, and an anchor test length of 40.

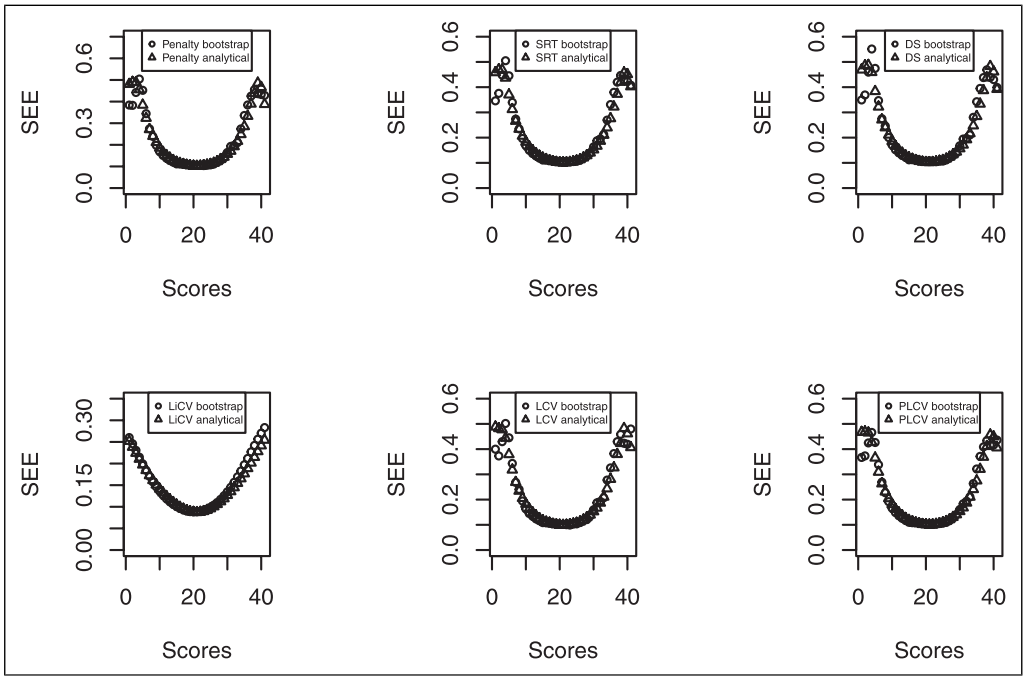

Figure 6. The analytical and bootstrap SEs for every KE estimator under the EG design. The results are based on the symmetric distributional scenario, 10,000 test takers per group and 40 items.

DS-based KE estimator also shows a similar pattern even though the analytical SE assumes that the bandwidth is a known constant.

Under the NEAT design with 1000 test takers, 80 items and 20 anchor items, the running times for calculating the bandwidth were 0.17 seconds (Penalty), 0.02 seconds (SRT), 0.03 seconds (DS), 5.13 minutes (LiCV), 0.20 seconds (LCV), and 0.28 seconds (PLCV). The relative performance were similar with 100 and 5000 test takers, and for the EG design. The LiCV clearly deviates because of its vast amount of calculations.

Empirical Illustration

The Swedish Scholastic Aptitude Test (SweSAT) is a large-scale standardized test used in the admission process to Swedish universities. It is a paper and pencil test that is given twice per year and consists of a quantitative and verbal section, each containing 80 items. The sections are equated separately. To illustrate the KE estimator for different bandwidth selection methods, we equated the quantitative section of the SweSAT using two consecutive administrations. The total sample consisted of 5609 test takers of which 2826 took the spring administration (test form X) and 2783 took the fall administration the year before (test form Y). The mean X score was 41.68 with standard deviation 32.46, and the mean Y score was 39.89 with standard deviation 29.16. The score distributions are both positively skewed, as can be seen in Figure 7. In practice, SweSAT employs the NEAT design since the assumptions underlying the EG design have been shown to be unfulfilled (Lyrén & Hambleton, 2011). This empirical illustration therefore uses the NEAT design as well. We applied post-stratification with a population weight reflecting the relative group sizes, and presmoothed the samples with log-linear models using the BIC measure to evaluate the goodness-of-fit. This resulted in models that preserved the first 4 moments of the marginal distributions of X/Y and A, respectively, and the first cross-moment of X/Y and A. The SEE and PRE for the first five moments were used to evaluate the equating results.

Empirical Illustration Results

In the upper part of Table 2, the resulting bandwidths for each method are presented. They clearly differ between each other, with the LCV selecting the smallest bandwidths and the SRT method

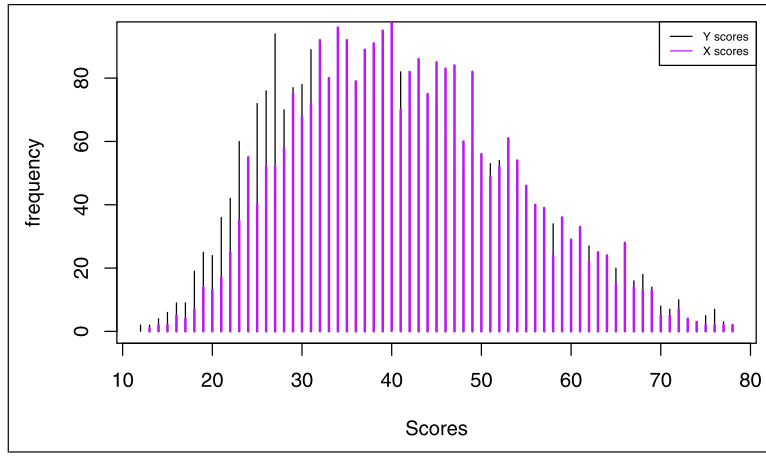

Figure 7. The score distributions for the SweSAT data, with the scores from the spring administration being represented as X scores, and the scores from the autumn administration on the previous year as Y scores.

Table 2. The selected bandwidths under the NEAT design together with the PRE of each KE estimator.

Bandwidth	Penalty	SRT	DS	LiCV	LCV	PLCV
h_X	0.73	2.25	0.75	1.75	0.34	0.67
h_Y	0.71	2.42	0.74	1.32	0.34	0.65
Moment						
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.001	−0.003	0.000
3	0.000	0.000	0.000	0.003	−0.009	0.000
4	0.000	0.001	0.000	0.006	−0.012	0.000
5	0.001	0.005	0.001	0.009	−0.027	0.001

the largest. The penalty method and the DS method selected very similar bandwidths, and the PLCV method selected bandwidths about twice the size of those selected by LCV. These findings are in line with those from the simulation study. By a graphical inspection, it was seen that the kernel density estimate using the LCV bandwidths resulted in a density with extensive fluctuations. The penalty function in the PLCV method has therefore been activated several times.

In the lower part of Table 2, the PRE of the five first moments are presented for all KE estimators. The PRE is small regardless of bandwidth selection method, but using the penalty, DS, and PLCV methods result in an equating transformation that best preserves the five first moment of the Y score distribution.

In the left panel of Figure 8, the difference between the equated scores and the raw scores are displayed for each KE estimator. The estimators produce very similar results over large parts of the score range, with visible differences only in the tails of the score scale.

In the right panel of Figure 8, the SEE of the KE estimators are shown. Again it is in the tails of the score scale where the differences are most evident. The SEE for all estimators is the largest in the tails, explained by the fact that there are fewer test takers with extreme scores. The PLCV-based estimator has some of the smallest SEEs for the top scores, and the LCV-based estimator has the highest peaks in the tails of the score range.

Discussion

The overall aim of this study was to compare different bandwidth selection methods in KE, since it is well known that the choice of bandwidth is an essential part of kernel density estimation (Sheather, 2004). Thus, it was important to investigate to what extent the bandwidth has an influence on the equating transformation and if that differed depending on the method used.

The results indicate that the KE estimator is, at least to some extent, insensitive to the choice of bandwidth. Although the selected bandwidths differ between the evaluated methods, the differences in the subsequent equating are small regardless of score distributional shape, sample size, and test length. However, the listed factors are affecting the equating error and variance for every evaluated method. These findings are in line with previous studies (Andersson & von Davier, 2014; Häggström & Wiberg, 2014; Liang & von Davier, 2014) which all found only small differences between their proposed methods and the penalty method. The differences seen in our study were particularly small in terms of MSE. This might be explained by the fact that the MSE is not able to compare the KE estimators over the whole score scale, but only on how well the first two moments of the equated scores correspond to the score distribution that the equated scores attempt to estimate. The results also showed that the MSE under the NEAT design were about half the size compared to that under the EG design, which is in line with the results of

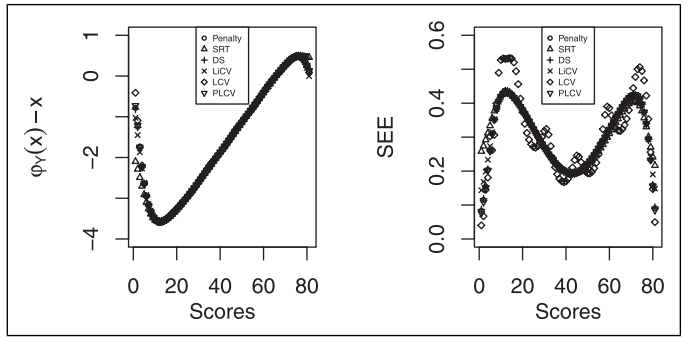

Figure 8. Left panel: The difference between the equated and raw scores for each KE estimator using the SweSAT data. Right panel: The SEE for each KE estimator using the SweSAT data.

Häggström and Wiberg (2014). We believe the reason for this is that the generated test groups were quite similar, although non-equivalent, as you would expect to see in a real testing situation. With an anchor that correlated strongly with the test scores, the variance of the equating transformation should decrease. Since the MSE for the most part was constituted by the variance, the MSE should thus be lower. We also compared every KE estimator with respect to the mean of the equated scores for every score point, and for the most part only small differences were found. However, the results showed that the bandwidth methods sometimes produced equated scores that were larger than a DTM.

The simulation results also showed that the analytical SEE got very close to the bootstrap SE regardless of bandwidth method but with a systematic error at the tails, the SRT method exempted. This means that the variability introduced by the bandwidth choice is not taken into account. For future research, it is thus of importance to derive accurate SEE formulas for all data-driven bandwidth methods.

In terms of PRE, the largest differences between the estimators were seen for the higher moments. At the same time, it should be noted that there are no clear guidelines from previous studies on how many higher moments that are meaningful to compare, or when the magnitude of the PRE indicates a poor equating estimate. Generally, the simulation study showed that shorter tests with more test takers result in smaller PREs and SEs, under both the EG and NEAT design. The promising performance of the LiCV method under the EG design seen in Liang and von Davier (2014) could thus not be repeated under the NEAT design. Moreover, the LiCV method took, by far, the longest time to compute which is not surprising since the procedure has to be repeated 1000 times. It is possible that the LiCV could be calculated using fewer iterations without losing its quality of performance, but this is left for future research.

In order to analyze the influence of bandwidth selection on KE, other factors such as log-linear model specification and the choice of kernel function were purposely marginalized. Since KE involves five steps that affect the equated scores, the simulation study cannot be viewed as exhaustive. One limitation is that we only investigated the impact of the bandwidth on the KE transformation using post-stratification equating. Although it would be of interest to examine the bandwidth impact using chained KE, we do not expect large differences since other studies have showed that post-stratification and chained KE usually give similar results. Future research should also investigate bandwidth selection for item response theory KE.

To conclude, the findings of this paper show that the choice of bandwidth in KE is not crucial in terms of equated scores, but that there still are factors that could make some of the bandwidth methods more appealing. The penalty, DS, and PLCV methods are most robust to changes in test

length, number of test takers, and score distributions. They are also quick to compute and could thus be recommended in practice. Our findings also makes the difference between equipercentile equating, linear equating and KE smaller, since the traditional approaches are part of KE as a special case. Practitioners can therefore make use of the flexibility of KE without having to be too concerned about the choice of smoothing parameter. Since previous research has reached similar conclusions regarding the choice of kernel function, the critical part of equating instead seems to lie at the first step, the log-linear presmoothing. The results of this study therefore gives further strength to the view of KE as a family of equating methods that both incorporates traditional and modern equating methods, rather than being a completely new method of equating. KE therefore offers an easy way to both equate test forms and perform sensitivity analysis of the results, by making it possible to compare not only a very smooth equating function or traditional equating, but every possible equating function in between these two modes.

Author's Note

Parts of the research was conducted while Gabriel Wallin was conducting postdoctoral research at: Université Côte dAzur, Inria, CNRS, Laboratoire J.A. Dieudonné, team Maasai.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Swedish Wallenberg grant MMW 2019.0129.

ORCID iDs

Gabriel Wallin https://orcid.org/0000-0002-7930-6701
Jenny Häggström https://orcid.org/0000-0002-9086-7403
Marie Wiberg https://orcid.org/0000-0001-5549-8262

Supplemental material

Supplemental material for this article is available online.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle Selected papers of hirotugu akaike. In: *Springer Series in Statistics*.(pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15.

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1–25. http://dx.doi.org/10.18637/jss.v055.i06.

Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement*, *51*(3), 223–238.

Braun, H., & Holland, P. (1982). Observed-score test equating: a mathematical analysis of some ets equating procedures. In P. Holland, & D. Rubin (Eds.), *Test equating* (Vol. 1, pp. 9–49). Academic Press.

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)*. ETS.

González, J., & von Davier, A. A. (2017). An illustration of the epanechnikov and adaptive continuization methods in kernel equating. In L. A. van der Ark, M. Wiberg, S. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 253–262). Springer International Publishing.

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.

Häggström, J., & De Luna, X. (2010). Estimating prediction error: Cross-validation vs. accumulated prediction error. *Communications in Statistics - Simulation and Computation*, *39*(5), 880–898. https://doi.org/10.1080/03610911003650409.

Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement*, *51*(2), 201–211. https://doi.org/10.1111/jedm.12042.

Hall, P., Marron, J. S., & Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, *92*(1), 1–20. https://doi.org/10.1007/BF01205233.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*(433), 401–407.

Lee, Y., & von Davier, A. (2011). Equating through alternative kernels. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (Vol. 1, pp. 159–173). Springer. http://dx.doi.org/10.1007/978-0-387-98138-3_10.

Liang, T., & von Davier, A. A. (2014). Cross-validation: An alternative bandwidth-selection method in Kernel equating. *Applied Psychological Measurement*, *38*(4), 281–295. http://dx.doi.org/10.1177/0146621613518094.

Lyrén, P.-E., & Hambleton, R. K. (2011). Consequences of violated equating assumptions under the equivalent groups design. *International Journal of Testing*, *11*(4), 308–323. https://doi.org/10.1080/15305058.2011.585535.

Moses, T., & Holland, P. W. (2009). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement*, *46*(2), 159–176. https://doi.org/10.1111/j.1745-3984.2009.00075.x.

Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 557–574. https://doi.org/10.1348/000711009x478580.

Park, B. U., & Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, *85*(409), 66-72.

R Core Team (2018). *R: A language and environment for statistical computing [computer software manual]*. R Core Team. https://www.R-project.org/.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scott, D. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.

Sheather, S. (2004). Density estimation. *Statistical Science*, *19*(4), 588–597. https://doi.org/10.1214/088342304000000297.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, *12*(4), 1285–1297. https://doi.org/10.1214/aos/1176346792.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*(4), 605–623.

von Davier, A. A., Holland, P., & Thayer, D. (2004). *The kernel method of test equating*. Springer.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, *21*(4), 1–21. http://dx.doi.org/10.18637/jss.v021.i04.