<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Journal of educational measurement*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

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>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-89689>

Optimal Bandwidth Selection in Observed-Score Kernel Equating

Jenny Häggström¹

Marie Wiberg²

¹PhD, Department of Statistics, USBE, Umeå University, Sweden

²Associate professor, Department of Statistics, USBE, Umeå University, Sweden

Corresponding author

Marie Wiberg
Department of Statistics, USBE
Umeå University
SE-901 87 Umeå
Sweden

+46 90 786 95 24

marie.wiberg@stat.umu.se

This paper has been published. Please cite as:

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

Optimal Bandwidth Selection in Observed-Score Kernel Equating

Abstract

The selection of bandwidth in kernel equating is important because it has a direct impact on the equated test scores. The aim of this paper is to examine the use of double smoothing when selecting bandwidths in kernel equating and to compare double smoothing with the commonly used penalty method. This comparison was made using both an equivalent groups design and a non-equivalent group with anchor test design. The performance of the methods was evaluated through simulation studies using both symmetric and skewed score distributions. In addition, the bandwidth selection methods were applied to real data from a college admissions test. The results show that the traditional penalty method works well although double smoothing is a viable alternative because it performs reasonably well compared to the traditional method.

Optimal Bandwidth Selection in Observed-Score Kernel Equating

Test equating is used to obtain scores on different test forms that are comparable. In observed-score kernel equating, the continuization of the observed distributions of discrete test scores is based on statistical principles and a Gaussian kernel has typically been used (von Davier, 2011; von Davier, Holland & Thayer, 2004). In traditional equipercenile equating, linear interpolation has been used for this continuization (Kolen & Brennan, 2004). The kernel equating function is affected by the selection of bandwidths in this continuization (von Davier et al., 2004), and thus ultimately affects the test takers' test scores.

There are many ways to select a bandwidth. One popular method is cross-validation (Rudemo, 1982; Bowman 1984), and a special case of it has been used with kernel equating and the equivalent groups (EG) design (Liang & von Davier, 2009). It is also common to minimize a penalty function (Härdle & Marron, 1985), which is the approach typically used in kernel equating (von Davier et al., 2004). Recently, Cid and von Davier (2009) proposed the use of adaptive and plug-in methods. Although several methods exist, there is still no best way to select optimal bandwidths in kernel equating. In other areas, double smoothing methods have resulted in smaller mean squared errors than other bandwidth selection methods when estimating certain parameters (Häggström & de Luna, 2011; Häggström, 2013). The aim of this paper is to examine the use of double smoothing when selecting the bandwidth in kernel equating and to compare it to the commonly used penalty method. Both EG design and non-equivalent groups with anchor test (NEAT) design were used with both simulations and real data. The methods were evaluated with respect to bias, mean squared error, percent relative error, and the standard error of equating.

Kernel Equating

The aim of kernel equating (von Davier et al. 2004) is to find an optimal equating transformation between a new test form Y and an old test form X where X and Y are the scores on tests X and Y . Assume that the test takers with scores X and Y are random samples from population P and Q of test takers. Let $r_j = \Pr\{X = x_j; T\}$ and $s_k = \Pr\{Y = y_k; T\}$, where $j = 1, \dots, J$ and $k = 1, \dots, K$, are the probabilities of scoring $X = x_j$ and $Y = y_k$, respectively, for a test taker randomly selected from a target population T . Kernel equating comprises five steps. The first step is *presmoothing*, where appropriate statistical tools are used to smooth the raw score distributions prior to equating them. In the second step, we derive the *estimated score probabilities* $\hat{\mathbf{r}} = (\hat{r}_1, \dots, \hat{r}_J)^T$ and $\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_K)^T$ of X and Y for the target population T from the estimated score distribution in the first step: $\hat{F}_X(x) = \hat{\Pr}(X \leq x; T)$ and $\hat{F}_Y(y) = \hat{\Pr}(Y \leq y; T)$, respectively. The smoothed (joint) raw score distributions are transformed via a design function

into marginal distributions for T . In the third step, *continuization of discrete distributions*, one ensures that the cumulative distribution functions of the discrete test scores are continuous. A Gaussian kernel is typically used, which for test scores on X is

$$F_T(x; h_X) = F_{h_X}(x) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X) \mu_{XT}}{h_X a_X}\right),$$

where $\Phi(z)$ is the standard normal distribution function, h_X is the bandwidth, $\mu_{XT} = \sum_j x_j r_j$, $a_X = \sqrt{\sigma_{XT}^2 / (\sigma_{XT}^2 + h_X^2)}$, and $\sigma_{XT}^2 = \sum_j (x_j - \mu_{XT})^2 r_j$ (von Davier et al., 2004). We want the derived density functions $f_T(x; h_X)$ and $f_T(y; h_Y)$ from $F_T(x; h_X)$ and $F_T(y; h_Y)$, respectively, to be smooth. To ensure that they are reasonably smooth, one needs to select optimal bandwidths h_X and h_Y . The fourth step is the *equating* from test form Y to X

$$\hat{e}_X(y) = e_X(y; \hat{r}, \hat{s}) = F_{h_X}^{-1}(F_{h_Y}(y; \hat{s}); \hat{r}) = \hat{F}_{h_X}^{-1}(\hat{F}_{h_Y}(y)). \quad (1)$$

In the fifth step, the asymptotic *standard error of equating* (SEE) is calculated from

$$SEE_X(y) = \hat{\sigma}_X(y) = \sqrt{\text{Var}(\hat{e}_X(y))}. \quad (2)$$

Bandwidth Selection Methods in Kernel Equating

The selection of bandwidths is important in kernel equating because it affects the resulting equating function. A large bandwidth (i.e. $h_X > 10\sigma_X$) results in a very smooth continuized cumulative distribution, and the equating function becomes close to linear equating. A small bandwidth (e.g. $h_X = 0.01$) results in a distribution that varies more intensely and can lead to undersmoothed data. A bandwidth of $h_X = 0.33$ results in an equating close to equipercentile equating using linear interpolation (von Davier et al., 2004).

Minimizing a Penalty Function

In kernel equating, a penalty function is traditionally minimized in order to find the bandwidths h_X and h_Y . This is done similarly for them. To find the bandwidth for h_X minimize

$$\text{PEN}(h_X) = \sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2 + \kappa \cdot \sum_j A_j, \quad (3)$$

where $\hat{f}_{h_X}(x_j)$ is the estimated density function of the continuous transformation of X , κ is a constant typically equal to 0 or 1 depending on whether the second penalty term is used. In this equation, $A_j = 0$ or $A_j = 1$ and the latter is true if

$$[(\hat{f}'_{h_X}(x_j - w) > 0) \cap (\hat{f}'_{h_X}(x_j + w) < 0)], \text{ or } [(\hat{f}'_{h_X}(x_j - w) < 0) \cup (\hat{f}'_{h_X}(x_j + w) > 0)],$$

where $\hat{f}'_{h_X}(x_j)$ is the derivative of $\hat{f}_{h_X}(x_j)$ and w is a constant typically chosen to be 0.25 (Lee & von Davier, 2011; von Davier, 2013). The first term in Equation 3 is a comparison of the

estimated relative frequencies and the continuized density function at each score value, thus it preserves the characteristics of the distribution but may undersmooth the data. The second term ensures the smoothness of the continuized distribution. In the later empirical study, $\kappa 0$ means that only the first term is used and $\kappa 1$ means that both terms are used.

Double Smoothing

The idea of double smoothing (DS) was first proposed by Hall, Marron, and Park (1992) in the context of nonparametric density estimations and has since been studied for use in e.g. simple nonparametric regression estimations (Härdle, Hall & Marron, 1992), semiparametric estimations of partial linear models (Häggström, 2013), and nonparametric estimations of average causal effects (Häggström & de Luna, 2011). In general, the main idea of DS is to select the bandwidth that minimizes an estimate of the mean squared error of the targeted estimator where the bias part is estimated using a pilot bandwidth. Here, we propose the use of a new DS procedure. First, start with a very smooth first estimate of the density function. This is accomplished by using a subjectively chosen large pilot bandwidth, g_X , and estimating f_{g_X} at the score values and the values halfway between them $\mathbf{x}^* = \{x_l^*\} = [x_1, x_1 + 0.5, x_2, \dots, x_J - 0.5, x_J]^T$, $l = 1, \dots, 2J - 1$. The resulting estimate, $\hat{f}_{g_X}$, is very smooth but it will not perfectly interpolate the estimated score probabilities, the $\hat{r}_l$'s. Second, we improve on this first estimate by estimating f_{h_X} at $\mathbf{x}^*$ using $\hat{f}_{g_X}$ at the actual score values, $\mathbf{x}$, instead of using the $\hat{r}_l$'s. This yields a DS estimate $\hat{f}_{h_X}^*$. Finally, we select the bandwidth value of h_X that minimizes the sum of the squared difference between the l^{th} DS estimate $\hat{f}_{h_X}^*(x_l^*)$ and $\hat{r}_l^*$, where

$$\hat{r}_l^* = \begin{cases} \frac{\hat{r}_{l+1}}{2}, & \text{if } l \text{ is odd} \\ \hat{f}_{h_X}^*(x_l^*), & \text{if } l \text{ is even} \end{cases} \quad (4)$$

Formally, the proposed DS criterion is defined as

$$\text{DS}(h_X) = \sum_{l=1}^{2J-1} (\hat{r}_l^* - \hat{f}_{h_X}^*(x_l^*))^2, \text{ where} \quad (5)$$

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

with $\phi(z)$ denoting the standard normal density function, and

$$\hat{f}_{g_X}(x) = \sum_{j=1}^J \hat{r}_j \phi\left(\frac{x - \hat{a}_X^{g_X} x_j - (1 - \hat{a}_X^{g_X}) \hat{\mu}_{XT}}{g_X \hat{a}_X^{g_X}}\right) \frac{1}{g_X \hat{a}_X^{g_X}} \text{ with } \hat{a}_X^{g_X} = \sqrt{\hat{\sigma}_{XT}^2 / (\hat{\sigma}_{XT}^2 + g_X^2)}.$$

The pilot estimation part will effectively prevent undersmoothing, and the final DS part is an attempt to preserve the characteristics of the estimated relative frequency distribution.

Empirical Study

Both real data and simulations (with both skewed and symmetric distributions) were used. Skewed distributions were included because it is common that test score distributions are skewed. For each type of distribution, one simulated data set was generated for the EG design and one was generated for the NEAT design. The X and Y test scores ranged from 0 to 20, and the external anchor test score A ranged from 0 to 10. For the EG design, given $\mathbf{r} = (r_1, \dots, r_J)^T$ and $\mathbf{s} = (s_1, \dots, s_K)^T$, 1000 instances of $\mathbf{X} = (X_1, \dots, X_N)$ and $\mathbf{Y} = (Y_1, \dots, Y_N)$ were generated from multinomial distributions for both of the sample sizes $N = 100$ and $N = 1000$. For the NEAT design, post-stratification was used. Given the vectorized probability matrices for populations P and Q , $\mathbf{v}(\mathbf{P})$ and $\mathbf{v}(\mathbf{Q})$, respectively, (von Davier et al, 2004, p. 35) that contain the joint probabilities of test scores $p_{jl} = \Pr(X = x_j, A = a_l | P)$ and $q_{kl} = \Pr(Y = y_k, A = a_l | Q)$, 1000 instances of $\mathbf{X}$ and $\mathbf{A} = (A_1, \dots, A_N)$ in P and $\mathbf{Y}$ and $\mathbf{A} = (A_1, \dots, A_N)$ in Q were generated from multinomial distributions of each of the sample sizes. Throughout, a weight between population P and Q of 0.5 was used in computing the score probabilities. For $\kappa=1$ we used $w = 0.25$ and for DS we used $g_X = 4$, which we deemed large enough because larger pilot bandwidths seemed to give more or less the same end result.

In the simulation using the symmetric setting, the true score probabilities for the EG design were the fitted score probabilities from chapter 7 of von Davier et al. (2004) and the true score probabilities for the NEAT design were the `simeq` data found in the `kequate` package. In the skewed setting, the true score probabilities were the fitted score probabilities based on score samples created by multiplying beta random numbers (with shape parameters 5 and 2) by the maximum score value. These were then rounded to integers, which is an approach used by Liang and von Davier (2009). We used the functions `rbeta` and `rMvdc` in the R package for the EG and NEAT designs, respectively. For both distributions settings, we used the same log-linear models as in von Davier et al. (2004) in the EG design, and the same models as in the NEAT example in the R package `kequate` for the NEAT design.

Real data from a college admissions test that is administered twice a year was used. The test consists of a verbal and a quantitative section – both having 80 items – that are equated separately. The groups of test takers who take each administration of this test are usually similar and this makes it possible to use an EG design. An equipercentile equating with the EG design was performed from the new test Y completed by 56,358 test takers to the old test X completed

by 40,431 test takers. Anchor tests A containing 40 items for each section were administered to a smaller group of test takers. The verbal and quantitative anchor tests were administered to 2,480 and 5,263 test takers, respectively, who took the old test X and to 14,172 and 6,465 test takers, respectively, who took the new test Y. An equipercentile equating with a NEAT design was used for comparisons. Descriptive statistics of the real and simulated data are given in Table 1.

Table 1

Descriptive statistics for the test forms. The mean scores, standard deviations (SD), skewness, kurtosis, and correlation (r_{XA}) between anchor test A and either test form X or Y.

Data: design	Test form	Mean	SD	Skewness	Kurtosis	r_{XA}
<i>Symmetric: EG</i>	X	10.82	3.81	-0.07	-0.30	
	Y	11.59	3.93	-0.06	-0.43	
<i>Symmetric: NEAT</i>	X	10.52	4.17	0.01	-0.28	
	Y	9.30	3.89	0.11	-0.24	
	A (pop. P)	4.97	2.62	0.06	-0.74	0.70
	A (pop. Q)	5.84	2.85	-0.25	-0.81	0.78
<i>Skewed: EG</i>	X	14.40	3.20	-0.59	0.39	
	Y	14.52	3.07	-0.60	0.50	
<i>Skewed: NEAT</i>	X	14.22	3.19	-0.59	-0.13	
	Y	14.17	3.17	-0.48	0.06	
	A (pop. P)	7.07	1.61	-0.55	0.04	0.71
	A (pop. Q)	7.16	1.58	-0.62	-0.01	0.74
<i>Real Verbal: EG</i>	X	41.60	13.50	0.31	-0.58	
	Y	39.75	12.51	0.33	-0.53	
<i>Real Verbal: NEAT</i>	X	43.85	13.98	0.22	-0.66	
	Y	40.78	12.77	0.26	-0.60	
	A (pop. P)	20.94	7.81	0.16	-0.77	0.87
	A (pop. Q)	19.98	7.55	0.30	-0.69	0.87
<i>Real Quant: EG</i>	X	37.91	13.43	0.47	-0.48	
	Y	37.05	12.10	0.51	-0.31	
<i>Real Quant: NEAT</i>	X	39.66	13.59	0.36	-0.58	
	Y	38.43	12.24	0.42	-0.41	
	A (pop. P)	18.40	6.55	0.36	-0.35	0.85
	A (pop. Q)	18.46	6.74	0.40	-0.38	0.84

Evaluation Criteria

The equatings were evaluated with the mean of selected bandwidths, mean squared error (MSE), percent relative error (PRE), and SEE as defined in Equation 2. The squared bias and variance of $\hat{\mu}_X$ are defined by

$$\hat{\text{Bias}}^2(\hat{\mu}_X) = \left[\frac{1}{1000} \sum_{i=1}^{1000} (\hat{\mu}_{X_i} - \mu_X) \right]^2 \text{ and} \quad (6)$$

$$\hat{\text{Var}}(\hat{\mu}_X) = \left[\frac{1}{1000-1} \sum_{i=1}^{1000} \left(\hat{\mu}_{X_i} - \frac{1}{1000} \sum_{i=1}^{1000} \hat{\mu}_{X_i} \right) \right]^2, \quad (7)$$

where $\hat{\mu}_{X_i} = \sum_{k=1}^K \hat{e}_{X_i}(y_k) s_k$, and MSE is the sum of Equations 6 and 7. Similar quantities regarding $\hat{\mu}_Y$, $\hat{\sigma}_X$, and $\hat{\sigma}_Y$ were also computed. Bias and SEE are used to determine if there is a systematic or a random error, respectively. The PRE is an index of how well the distribution of the equated scores of the new test form Y matches the distribution of the old test form X to which the equating was intended. Let the p^{th} moments of scores from X and $e_X(Y)$ be denoted by $\mu_p(X) = \sum_{j=1}^J (x_j)^p r_j$ and $\mu_p(e_X(Y)) = \sum_{k=1}^K (e_X(y_k))^p s_k$, then the PRE is defined as

$$\text{PRE}(p) = 100 \frac{\mu_p(e_X(Y)) - \mu_p(X)}{\mu_p(X)}, \quad (8)$$

(von Davier et al., 2004). The evaluation and simulations were performed using R (R Core Team, 2013) and the kernel equating package kequate (Andersson et al., 2013).

Results

Results based on the simulated data are summarized in the left panel of Table 2. Bias and MSE for $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ were similar and low for all methods, thus only results are reported for $\hat{\mu}_X$. For the NEAT design, DS resulted in bandwidth values that were on average larger than bandwidths selected by $\kappa 0$ but smaller than the values selected by $\kappa 1$ in the symmetric settings. In the skewed setting, $\kappa 0$ and DS resulted in similar bandwidths. For the EG design with the symmetric setting, DS selected values of h_Y that were, on average, in between those obtained with the penalty methods, and the same pattern was observed for the NEAT design. This was also true for h_X in the EG design with the skewed setting. For the EG design with the symmetric setting, DS resulted in slightly larger values on average for h_X than the other methods. For the EG design with the skewed setting, DS resulted in slightly smaller h_Y values on average than the other methods.

The PRE gave similar results across settings, thus only the highest average PRE value is presented. For both the symmetric and skewed distributions, DS yielded generally low PRE values that were smaller than or only slightly larger than those obtained with $\kappa 0$. These values were considerably less than those obtained with $\kappa 1$ in both the NEAT and EG designs. The method with the lowest PRE in all conditions except the EG design with skewed data was $\kappa 0$. $\kappa 1$ performed in general worst with respect to PRE. The average SEEs for the three conditions and the selected bandwidths were plotted and can be obtained upon request from the authors. In general, the SEE values were similar regardless of method, although $\kappa 1$ had lower average SEE values at more score values than the other methods in all settings except for the NEAT design with skewed data where DS had a lower average $\text{SEE}_Y(x)$ at more score values. $\kappa 0$ resulted in the highest average SEEs at more score values than the other methods in the symmetric setting, but in the skewed setting $\kappa 1$ had higher average $\text{SEE}_Y(x)$ and DS had

higher average $SEE_x(y)$. Although, the mean bandwidths differed somewhat, the difference between $\kappa 0$ and DS were negligible in the EG design except in the skewed setting where there were larger differences at the endpoints, especially the lower ones. Similarly, in the NEAT design $\kappa 0$ differed only slightly compared with DS, although the differences were somewhat larger at the endpoints with a symmetric distribution. $\kappa 1$ resulted in the largest equated score differences compared with the other methods, especially at the endpoints.

Table 2

The left side of the table shows results from the simulations. The mean bandwidths (h_X and h_Y) and variances (var), MSE, the squared bias when estimating μ_X , and the highest average PRE value over replicates. The upper right part of the table shows the selected bandwidths (h_X and h_Y) and the highest PRE value for the real data. The lower right part shows the equated test scores for the verbal section for a selection of test scores.

Simulated data							Real data						
NEAT design				EG design			NEAT design				EG design		
$\kappa 0$		$\kappa 1$	DS	$\kappa 0$	$\kappa 1$	DS	$\kappa 0$		$\kappa 1$	DS	$\kappa 0$	$\kappa 1$	DS
<i>Symmetric</i>							<i>Quantitative</i>						
$N = 1000$							h_X	0.650	2.422	0.589	0.633	0.710	0.570
h_X	0.612	0.679	0.648	0.622	0.640	0.662	h_Y	0.646	1.803	0.583	0.637	1.632	0.572
Var	6e-06	0.061	8e-05	7e-05	0.015	8e-05	PRE	0.070	0.650	0.052	0.060	0.450	0.050
h_Y	0.619	0.689	0.659	0.571	0.814	0.607							
Var	5e-05	0.056	6e-05	1e-04	0.188	1e-04	<i>Verbal</i>						
MSE	0.016	0.016	0.014	0.029	0.029	0.029	h_X	0.655	1.048	0.621	0.661	2.407	0.616
Bias ²	4e-04	4e-04	5e-04	3e-06	4e-06	4e-06	h_Y	0.659	0.734	0.615	0.657	0.767	0.611
PRE	0.98	1.31	1.08	-1.76	-2.22	-1.37	PRE	0.060	0.150	0.056	0.030	0.430	0.027
$N = 100$													
h_X	0.609	0.726	0.644	0.626	0.647	0.660	<i>Verbal test scores</i>						
Var	5e-04	0.109	7e-04	8e-04	0.013	7e-04	score	$\kappa 0$	$\kappa 1$	DS	$\kappa 0$	$\kappa 1$	DS
h_Y	0.615	0.733	0.655	0.569	0.846	0.593							
Var	4e-04	0.114	6e-04	9e-04	0.239	8e-04	0	-0.46	-0.85	-0.45	-0.80	-5.25	-0.79
MSE	0.138	0.138	0.138	0.309	0.310	0.309	1	0.35	0.12	0.35	0.15	-3.78	0.15
Bias ²	3e-04	2e-04	2e-04	9e-05	9e-05	9e-05	2	1.29	1.11	1.29	1.19	-2.35	1.19
PRE	0.90	1.47	0.99	-2.49	-3.03	-2.53	3	2.33	2.16	2.32	2.25	-0.94	2.24
<i>Skewed</i>							4	3.40	3.24	3.39	3.31	0.45	3.30
$N = 1000$							5	4.48	4.34	4.48	4.37	1.81	4.36
h_X	0.555	1.343	0.557	0.533	1.044	0.548							
Var	3e-05	0.309	6e-05	4e-05	0.453	6e-05	10	9.95	9.87	9.95	9.66	8.31	9.66
h_Y	0.562	1.293	0.569	0.559	0.705	0.535	20	20.94	20.93	20.94	20.23	20.06	20.23
Var	4e-05	0.252	7e-05	5e-05	0.192	1e-04	30	32.00	32.00	32.00	30.98	31.06	30.98
MSE	0.010	0.010	0.009	0.017	0.017	0.018	40	43.09	43.09	43.09	41.96	42.01	41.96
Bias ²	8e-06	3e-06	4e-05	2e-04	1e-04	1e-05	50	54.08	54.07	54.08	52.79	52.74	52.79
PRE	0.04	0.73	0.14	-0.08	1.40	-0.04	60	64.88	64.87	64.88	63.32	63.21	63.32
$N = 100$							70	75.18	75.19	75.18	73.93	74.08	73.93
h_X	0.557	0.943	0.558	0.536	0.764	0.549							
Var	3e-04	0.316	0.558	4e-04	0.212	7e-04	75	79.08	79.14	79.08	78.82	79.60	78.82
h_Y	0.564	0.966	0.569	0.554	0.851	0.540	76	79.58	79.69	79.58	79.54	80.64	79.53
Var	3e-04	0.327	6e-04	5e-04	0.255	0.013	77	80.00	80.19	79.98	80.10	81.67	80.09
MSE	0.106	0.107	0.092	0.204	0.205	0.204	78	80.35	80.67	80.33	80.56	82.70	80.53
Bias ²	3e-07	2e-06	1e-04	2e-05	6e-06	1e-05	79	80.70	81.16	80.67	80.97	83.76	80.91
PRE	-0.01	0.58	0.07	-0.12	-0.08	-0.06	80	81.11	81.75	81.06	81.39	84.95	81.31

The selected bandwidths, the highest PRE, and the equated scores obtained from the real data are shown in the right panel of Table 2. DS gave considerably smaller bandwidths than the penalty methods. Although $\kappa 0$ and DS gave different bandwidths their bandwidths did not

change across data collection design. PRE was in general low for all methods regardless of the data collection design, and the highest value observed was 0.65 ($\kappa 1$). The lower right part of Table 2 contains the equated values for every tenth value together with the endpoint values for the verbal section for all methods using both designs. All methods yielded similar equated scores in the NEAT design. Regardless of bandwidth selection method used, the equated scores in the mid part of the test score range are about the same, but they differ at the endpoints in the EG design. The large differences at the endpoints in the EG design for $\kappa 1$ compared to the other methods, especially at the higher end of the scores, could have serious consequences for test takers because they use their test scores for college admission. Another interesting result is the apparent differences in equated test scores between the EG and the NEAT designs. In the EG design, we could use all test takers who took the test even though we lacked information from common items. In the NEAT design, only a limited number of test takers could be used because only some of the test takers were given the anchor test due to test security concerns. Thus we have less information of the test takers, and this is also illustrated by the differences in the descriptive statistics in Table 1. The results for the quantitative section were similar to those for the verbal section shown in Table 2.

In general, the SEE values were lower for the EG design for all methods compared with the NEAT design for both the test sections. All three methods gave almost identical SEE curves within a given design. $\kappa 1$ had lower $SEE_Y(x)$ while DS had lower $SEE_X(y)$ at more score values than the other methods using both the EG and NEAT designs. However, $\kappa 1$ and DS also had higher $SEE_Y(x)$ and $SEE_X(y)$, respectively, at more score values than the other methods for both the EG and NEAT designs.

Concluding Remarks

This paper examined the possibility of using DS to select bandwidths in kernel equating instead of using the traditional penalty method. The DS criterion differs from the penalty method in terms of how deviation from smoothness is penalized. The performance of the DS method resulted in slightly different bandwidths than the traditional penalty method, but it did not result in qualitatively different equatings except at the endpoints. The PRE in the simulations and the real data study was, in general, low for all methods, and the SEE values were similar between the methods. One might argue that the differences are small, but because there are practical implications of equating test scores it is important to compare bandwidth selection methods. In general, the choice of bandwidth is considered to be more crucial in e.g. kernel regression than the choice of kernel (Wasserman, 2006). This was not seen in this study. Instead, the observed equated score differences in the middle of the score scale in Table 2 are essentially similar to the size differences in the study of Lee and von Davier (2011) who

examined differences in equated test scores when using different kernels. However, there were large differences at the endpoints for κ_1 with the EG design compared with the other methods in the real data study.

From the real data study, it was evident that the choice of bandwidth selection had an impact on the test scores at the endpoints. Because the test takers use the test scores for college admission, the scores at the upper end points are of great importance. An interesting result was that there was a generally large difference for all equated test scores between the NEAT and EG designs. In a large-scale assessment, the anchor test might not be distributed to all test takers due to test security concerns and this means that fewer test takers can be part of the NEAT design than the EG design. Although the example may be viewed as simplistic, it triggers the question of which data collection design is the most informative. Do we get a more precise measure if we use the EG design because we can use a greater portion of test takers than we get if we use fewer test takers, but common items and thus have more information per test taker, as in the NEAT design?

One limitation was that we only varied the number of test takers even though it is also possible to vary the test length. The impact on SEE, MSE and bias when varying the test length or other conditions should be examined further in future studies. An apparent challenge is to examine which criteria should be used when selecting optimal bandwidths. In this study we used bias, MSE, PRE, and SEE. Bias and MSE are not satisfying criteria when we have real data and the true parameters are unknown. PRE gives too little information on which bandwidth to choose, so we are left with the SEE. In contrast to κ_1 , both κ_0 and DS gave almost the same bandwidths across sample sizes, and this makes these methods more consistent, or insensitive, depending on one's perspective. It is crucial that a bandwidth selection method fulfills all of these criteria, which DS does. The conclusion is that the penalty method works well although DS is a viable alternative because it performed reasonably well and is easy to implement.

References

- Andersson, B, Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating using the R package kequate. *Journal of Statistical Software* 55, 1-25.
- Bowman, A.W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika* 71, 353-360.
- Cid, J. & von Davier, A. A. (2009). Examining potential boundary bias effect in kernel smoothing on equating. Paper presented at NCME, 14-16 April, 2009, San Diego, CA.
- Häggström, J. (2013). Bandwidth selection for backfitting estimation of semiparametric additive models: A simulation study. *Computational Statistics and Data Analysis* 62, 136-148.

- Häggström, J. & de Luna, X. (2011). Data-driven smoothing parameter selection for estimating average treatment effects. In *Selection of smoothing parameters with applications in causal inference*, PhD thesis. Department of Statistics, Umeå University.
- Härdle, W., Hall, P. & Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association* 87, 227-233.
- Härdle, W. & Marron, J. S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* 72, 481-484.
- Hall, P., Marron, J. S. & Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields* 92, 1-20.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: methods and practices*. (2nd ed.). New York: Springer.
- Liang, T., & von Davier, A. A. (2009). Cross validation: An alternative continuization method in kernel equating. Paper presented at the International meeting of Psychometric Society, July 20-24, Cambridge, UK.
- Lee, Y.-H. & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking*. Chapter 10: pp. 159-173. New York: Springer.
- Ramsey, J.O., & Silverman, B.W. (2006). *Functional data analysis*. (2nd ed.). New York: Springer.
- R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics* 9, 65-78.
- von Davier, A. A. (2013). Observed-score equating: an overview. *Psychometrika* 78, 605-623.
- von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking*. Chapter 1: pp. 1-17. New York: Springer.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer-Verlag.
- Wasserman, L. (2006). *All of nonparametric statistics*. New York: Springer-Verlag.