<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):

Wiberg, M., van der Linden, W J., von Davier, A A. (2014)
Local Observed-Score Kernel Equating
Journal of educational measurement, 51(1): 57-74
<https://doi.org/10.1111/jedm.12034>

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-85135>

Running head: *Local kernel equating*

Local Observed-Score Kernel Equating

Marie Wiberg¹

Wim J. van der Linden²

Alina A von Davier³

¹Associate professor, Department of Statistics, Umeå University, Sweden

² Chief Research Scientist, CTB/McGraw-Hill, Monterey, CA, USA

³ Strategic Advisor, Educational Testing Service, Princeton, New Jersey, USA

Acknowledgement: The authors are grateful to René Lawless for her editorial comments on the previous version of the paper. The opinions expressed in this paper are of the authors and not necessarily of Educational Testing Service.

*Requests for reprints should be addressed to Marie Wiberg, Department of Statistics, USBE, Umeå University, SE-901 87 Umeå, Sweden (e-mail: marie.wiberg@stat.umu.se).

This paper has been published. Please cite as:

Wiberg, M., van der Linden, W.J., & von Davier, A. (2014). Local kernel observed-score equating. *Journal of Educational Measurement*, 51(1), 57-74.
<https://doi.org/10.1111/jedm.12034>

Local Observed-Score Kernel Equating

Abstract

Three local observed-score kernel equating methods, integrating methods from the local equating and kernel equating frameworks, are proposed. The new methods were compared with their earlier counterparts with respect to such measures as bias, as defined by Lord's (1980) criterion of equity, and percent relative error. The local kernel IRT observed score equating method, which can be used for any of the common equating designs, had small amount of bias, low percent relative error, and relatively low kernel standard error of equating, even when the accuracy of the test was reduced. The local kernel equating methods for the nonequivalent groups with anchor test had generally low bias and were quite stable against changes in the accuracy or length of the anchor test. Although all proposed methods showed small percent relative errors, the local kernel equating methods for the nonequivalent groups with anchor test design had somewhat larger standard error of equating than their kernel method counterparts.

Introduction

Testing programs typically administer multiple test forms, each on a different occasion. These test forms are constructed according to the same specifications and are intended to be parallel in terms of difficulty, content, display, etc. in order to be fair to the test takers who take them. As such, test forms seldom appear to be exactly parallel, and an additional statistical adjustment, such as score equating, is necessary to obtain scores on different forms that are entirely comparable.

Over the past decade, two new frameworks for observed-score equating have been proposed. One framework is the kernel method of observed-score equating (von Davier, Holland, & Thayer, 2004a; 2004b; von Davier, 2011; von Davier, 2013); the other is the local equating method (van der Linden, 2000; 2006a; 2006b; 2011; van der Linden & Wiberg,

2010; Wiberg & van der Linden, 2011). The two frameworks are motivated by different goals and are organized around entirely different principles.

Kernel equating was proposed as an equating method (von Davier, Holland & Thayer, 2004) but, due to its large implications (von Davier, 2011; 2013), can nowadays be considered an observed-score equating framework. Kernel equating is motivated by the idea that traditional observed-score equating methods developed for different equating designs should be treated in a unified approach and the continuization of the test score distributions should be based on statistical principles (von Davier, 2011; von Davier et al., 2004a). Local equating has been driven by the wish to approximate Lord's (1980) equity criterion of equating as closely as possible for scores on test forms that are not necessarily perfectly reliable or strictly parallel (van der Linden, 2000, 2011).

In spite of these more fundamental differences, both methods use the same core equation to equate scores, albeit for different types of distributions. Assume that we have two Test Forms, X and Y , with scores of X and Y , and we want to equate Y to X . For convenience, suppose that X and Y are continuous. We want to perform the equating on two different distributions of X and Y , which have (strictly increasing) cumulative density functions (cdf's) $F_X(x)$ and $F_Y(y)$. The core equation is the equipercentile equating transformation function from Y to X defined as:

$$x = \phi_X(y) = F_X^{-1}(F_Y(y)). \quad (1)$$

Assume that Test Form X is administered to a sample of test takers denoted here by P and Test Form Y is administered to a sample, Q . In traditional and kernel equating, Equation 1 is defined on a target population T . If the two groups of test takers that take X and Y , respectively, are either the same sample or are randomly drawn from the same target population T , then $T = P = Q$. If the two samples of test takers, P and Q , are from different populations, say also denoted by P and Q , then the distribution function for the target

population is defined as a weighted average of the two distribution functions for P and Q , with weights w and $1-w$, $0 < w < 1$, respectively. T defined as such is often called a synthetic population. If the samples of test takers are from different populations, then a common set of items called an anchor test taken by samples is used to adjust the differences in ability in the two populations. Moreover, in order to obtain the cdf's F_X and F_Y in Equation 1 on T , additional assumptions are made about the missing data, that is, about the test takers from P who are not administered Test Form Y and about the test takers from Q who are not administered Test Form X. As the actual procedure may involve the sampling of the test takers, equating error in the kernel framework always equals the sampling error.

Local equating does not start with the idea of combining two different populations of test takers into a new population before the equating. Instead, for each ability level measured by X and Y , the observed score distributions of all test takers at this level are used to apply the transformation in Equation 1. As the actual procedure involves the use of fallible information to identify the ability level of each test taker, the primary type of error addressed in local equating framework is measurement error.

Another way of viewing the two different types of distributions used in Equation 1 by the two different methods is that kernel equating always is performed on estimates of marginal distributions of X and Y for a (synthetic) population of test takers, whereas local equating is performed on the conditional distributions of X and Y given an estimate of (a monotone function of) the individual test taker's ability level.

The goal of this paper is to combine the formal tools of kernel equating (presmoothing; estimation of score probabilities; continuization of score distributions) with the types of distributions addressed in local equating. The results will be compared to (i) kernel equating for the distributions typically addressed in the kernel equating framework and (ii) local equating without the use of any of the kernel equating tools. A comparison will be

made for the nonequivalent-groups-with-anchor-test (NEAT) design in combination with chain equating, post-stratification, local equating (LE), and item response theory (IRT) observed-score equating. The results will be evaluated using such statistical criteria as bias, root mean squared error, percent relative error, and standard errors of equating.

It is hypothesized that by combining the two frameworks one can address both sources of errors – measurement and sampling error. For example, the data available to estimate the conditional observed-score distributions may be sparse for some of the ability levels, and one might improve the precision of these estimates using presmoothing. Similarly, by using a kernel equated score for distinct ability levels, one can account for the differences in ability between the test takers on the two equated test forms. The most important contribution of this integrated framework is perhaps conceptual. It illustrates the relationship among the various equating methods, sheds a different light on the definition of score comparability, and might aid to the understanding of the synergies between observed test scores, ability estimates, measurement, and sampling errors when the scores are made comparable.

The rest of this paper is structured as follows. In the next two sections we review the main principles of local and kernel equating. The fourth and fifth sections describe the equating methods used in this paper as well as the setup of our empirical study. Our results are presented and discussed in the last two sections.

Local Equating

As already indicated, local equating has emerged from Lord's (1980) definition of equity (van der Linden, 2000; van der Linden, 2006a, b; van der Linden & Wiberg, 2010). Informally, equity is achieved when the distribution of equated scores is indistinguishable from the scores to which they are equated. van der Linden (2000) demonstrated that this equity criterion implies that a family of equating transformations exists – with the use of one of its transformations for each ability level. Traditionally all equating methods have used

single equating transformations for entire populations of test takers. From the perspective of local equating, the use of a single transformation constitutes a compromise between the different transformations actually required for each ability level. Hence, statistically, the use of a single equating transformation will always lead to equatings that are biased and population dependent.

Assume that Test Forms X and Y measuring the same ability, θ . If we knew the value of θ for a test taker, the necessary equating transformation would be an application of Equation 1 to the conditional distributions of X and Y , given θ . Hence, for any number of test takers, the family of true transformations is:

$$x = \varphi^*(y; \theta) = F_{X|\theta}^{-1}(F_{Y|\theta}(y)), \quad \theta \in \mathfrak{R}, \quad (2)$$

with θ indexing the individual members of the family (van der Linden, 2000). To perform these equating transformations, the scores distributions need to be continuous, and in local equating, linear interpolation has traditionally been used to continuize them. In real-world applications, we do not know the students' true ability levels and have to use a statistical estimate or proxy (for the definition of a proxy, see below). An example of the former is the maximum likelihood estimate or Bayesian estimate of θ for an item response theory (IRT) model fitted to the response data. An example of the use of a proxy is the anchor test scores in the NEAT design (van der Linden, 2011; van der Linden & Wiberg, 2010). Of course, the use of both types of quantities involves measurement error. But, as already noted, this is the primary type of error addressed in local equating.

Kernel Equating

Kernel equating (KE) (von Davier et al., 2004a) is a unified framework to observed-score equating that comprises five different steps: presmoothing; estimation of the score probabilities; continuization of discrete distributions; equating; and calculating the standard

error of equating. These steps are briefly described below. The overall aim of KE is to find an optimal version of the equating transformation in (1) between X and Y for a target population T . In order to be successful, we need to obtain the cdf's of X and Y on T , i.e.,

$$F_X(x_j) = Pr(X \leq x_j; T) \text{ and } F_Y(y_k) = Pr(Y \leq y_k; T).$$

The five steps involved in KE are:

- 1) *Presmoothing* This step aims to use an appropriate statistical tool to smooth the raw score distributions for equating. It is a purely statistical step, where the tool that produces distributions that best fit the data should be our favorite.
- 2) *Estimation of the Score Probabilities*. In this step, the score probabilities r and s of X and Y for the target population T are obtained from the estimated score distribution in the first step: $F_X(x_j) = Pr(X \leq x_j; T)$ and $F_Y(y_k) = Pr(Y \leq y_k; T)$, respectively. Define $r_j = Pr\{X = x_j; T\}$ and $s_k = Pr\{Y = y_k; T\}$ as the score probabilities for a test taker randomly selected from the target population T scoring $X = x_j$ and $Y = y_j$, respectively.

An important tool used in this step is the design function that transforms the smoothed (joint) raw score distributions into marginal distributions for T . The specific shape of the design function depends on the design of the equating study.

- 3) *Continuization of Discrete Distributions*: Since test scores are discrete in their nature, we need to make their cdf's continuous to ensure the existence of Equation 1. Traditionally, linear interpolation has been used to continuize the score distributions (for instance, see Kolen & Brennan, 2004). Instead, KE uses a Gaussian kernel to make the test score distribution continuous which, for X is defined as:

$$F_T(x; h_X) = \sum_j r_j \Phi \left(\frac{x - a_X x_j - (1 - a_X) \mu_{XT}}{h_X a_X} \right) \quad (3)$$

where h_X denotes the bandwidth, $\Phi(z)$ denotes the standard normal distribution function or Gaussian distribution,

$$\mu_{XT} = \sum_j x_j r_j, \sigma_{XT}^2 = \sum_j (x_j - \mu_{XT})^2 r_j \text{ and } a_x = \sqrt{\sigma_{XT}^2 / (\sigma_{XT}^2 + h_X^2)} \quad (4)$$

(von Davier et al., 2004a). Continuization of the Y distribution is obtained similarly. In order to compromise between having the continuous distribution being as close as possible to the score probabilities r_j and being as smooth as possible without smoothing away any of the fundamental characteristics of the discrete distributions, the use of penalty functions for finding the optimal bandwidths h_X is recommended. For instance, the bandwidth h_X is found by minimizing

$$\text{PEN}(h_X) = \sum_j (\hat{r}_j - \hat{F}'_T(x, h_X))^2 + \kappa \sum_j B_j, \quad (5)$$

where κ is a constant between 0 and 1 depending if the second penalty is used. Let

$$\hat{F}'_T(x, h_X) = \hat{f}'_{h_X}(x) \text{ and further on, } B_j = 1 \text{ if}$$

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

where w_{pen} is a constant typically chosen to be .25. Otherwise B_j is equal to 0. In other words, the second penalty adds a constant whenever the derivatives of the continuized density function have different signs when evaluated at a small distance around each score value. This means that the role of the second penalty function is to smooth irregularities in the continuous distribution. This step is similar for s and Y . Note that KE thus also leads to a family of equating functions but with an entirely different purpose. The family depends on the bandwidths h_X and h_Y , which are chosen to be optimal for population distributions, not for individual test takers (von Davier, 2011).

- 4) *Equating*. This step involves deriving the actual equating transformation from Y to X between the two continuized cdf's obtained in the previous steps as:

$$x = \hat{\varphi}_X(y) = \varphi_X(y; \hat{r}, \hat{s}) = F_{h_X}^{-1}(F_{h_Y}(y; \hat{s}); \hat{r}) = \hat{F}_{h_X}^{-1}(\hat{F}_{h_Y}(y)) \quad (6)$$

- 5) *Calculating Standard Error of Equating*. One of the advantages of KE is that it provides a neat way to compute an asymptotic standard error of equating for random sampling from the target population. The kernel standard error of equating (SEE) is defined in general as:

$$SEE_X(y) = \hat{\sigma}_X(y) = \sqrt{\text{Var}(\varphi_X(y))}, \quad (7)$$

where $\hat{\varphi}_X(y)$ is defined in Equation 6. It is important to note, that the SEE depends upon the design functions associated with the different equating designs and upon the model used to presmooth the data.

Local Kernel Equating Methods

Local Kernel Equating for Any Design

Our first method is for use with any equating design provided that the two test forms jointly fit an item response theory (IRT) model. We call this method *local kernel IRT observed-score equating*. The method is rooted in IRT observed-score equating (Kolen & Brennan, 2004; van der Linden, 2011) and local equating conditional on ability (van der Linden, 2000) further referred to as local equating with estimated ability, as well as kernel IRT observed-score equating (von Davier, 2010). It is important that the IRT model fits Tests X and Y jointly - a condition satisfying the general requirement that the two tests that are equated should measure the same variable. The observed number-correct scores given the ability θ are compound binomial distributions, and we can use Lord and Wingersky's (1984) algorithm to generate them for Test Forms X and Y . Once these conditional distributions are

continuized, we can derive a family of equating transformations from these distributions of X and Y given θ :

$$x = \varphi^{L\theta K}(y) = \tilde{F}_{X|\hat{\theta}}^{-1}(\tilde{F}_{Y|\hat{\theta}}(y)), \theta \in \mathfrak{R} \quad (8)$$

where $\tilde{F}$ represents a continuized distribution. In the actual equating, we estimate θ from one test form for all test takers and use a member of the family of transformations from Equation 8 to equate the observed score Y to the score on X . The design function will be the identity function, since we are using equivalent test takers, i.e., test takers with the same level of ability:

$$r_j = \Pr(X = x_j | \theta) \quad \text{and} \quad s_k = \Pr(X = y_k | \theta). \quad (9)$$

Local Kernel Equating for the NEAT Design

Three different methods for the NEAT design are proposed. The methods differ in the way they handle the information in the anchor test. In the first method, we can apply the same *local kernel IRT observed-score equating method* as above, estimating θ from the data from Test Forms X and Y and Anchor A , produced by the sampling design. The samples of test takers from P and Q do not need to be randomly equivalent, i.e. they can be nonequivalent as defined in e.g. Kolen & Brennan (2004). The estimation of θ can be done either concurrently (but with two different ability distributions for the two groups in the case of marginal maximum likelihood), or separately from the data for Test Form X and Anchor Test A taken by group P and the data from Test Form Y and Anchor Test A taken by group Q , with a subsequent linking of the parameters using, for instance, the Stocking and Lord (1983) method. All other steps are exactly as described previously.

The second proposed method, the *local kernel equating conditional on anchor test score* method, has its roots in local equating conditional on an anchor-test score (van der Linden and Wiberg, 2010). This method relies on the assumption that the observed score on

Anchor Test A, is an accurate proxy of the common proficiency measured by Test Forms X and Y, where a proxy is defined as a monotone function of the proficiency. As shown by the test-characteristic function, A is always an accurate proxy in this sense provided the anchor test is long enough. Hence, except for measurement, the conditional distributions of X and Y given Anchor Score $A=a$ are the same as those given the proficiency. Once these distributions have been presmoothed and continuized, we obtain the family of equating transformations:

$$x = \phi^{LAK}(y) = \hat{F}_{X|a}^{-1}(\hat{F}_{Y|a}(y)), \quad a \in A \quad (10)$$

where $\hat{F}$ represents a presmoothed and continuized conditional distribution. Again, because we are using equivalent test takers (i.e., different test takers with the same anchor test scores),

$$r_j = \Pr(X = x_j | A) = \sum_k p_{jk} \quad \text{and} \quad s_k = \Pr(X = y_k | A) = \sum_j p_{jk}, \quad (11)$$

the design function used during presmoothing and continuization is the identity function. The KE counterparts of the proposed method are chain kernel equating (CK) and post-stratification kernel equating (PK) (von Davier, et al., 2004a).

The third proposed method is the *local kernel equating with ability estimated from the anchor test* method, which has its roots in local equating conditional on ability (van der Linden, 2000) and local equating with ability estimated from the anchor-test scores (Janssen, Magis, San Martin & Del Pino, 2009). There are no KE counterparts for this method. The method uses the same earlier assumption of the IRT model fitting the three tests jointly. As before, the conditional distributions are continuized and the family of equating transformations is derived from the distributions of X and Y, given θ :

$$x = \phi^{L\theta AK}(y) = \tilde{F}_{X|\theta}^{-1}(\tilde{F}_{Y|\theta}(y)), \quad \theta \in \mathfrak{R}, \quad (12)$$

In this case, θ is estimated from the anchor test. Again, the design function is the identity function:

$$r_j = \Pr(X = x_j | \theta) \quad \text{and} \quad s_k = \Pr(X = y_k | \theta) \quad (13)$$

Empirical Study

The goal of this study was to empirically evaluate the three proposed local kernel methods relative to their local and kernel methods counterparts, as shown in Table 1. Since we wanted to evaluate these methods for bias, the use of computer simulations allowed us to compare the estimated equating transformations with the true transformations defined in Equation 2. The test scores on the test forms were simulated under the 3-parameter logistic (3PL) IRT model. Although the choice for this model restricts the generalizability of our conclusions, it has two distinct advantages. First, the two tests need to measure the same construct (i.e., fulfill the assumption of unidimensionality), and the 3PL model is the most commonly used unidimensional model throughout the testing industry. Second, we wanted to be able to compare the results with those in earlier studies by van der Linden and Wiberg (2010) and Wiberg and van der Linden (2011).

The main factor varied in this study was the accuracy of the anchor-test scores (i.e., the length of the anchor test and the discrimination parameters of its items). In previous local equating research (van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011), this factor emerged as the most important.

The study was performed using Matlab and the freely available R package for kernel equating, *kequate* (Andersson, Branberg, & Wiberg, 2013a;b).

Evaluation Criteria

As the choice of only one evaluation criterion may favor one method over another (Sinharay & Holland, 2008), multiple evaluation criteria from both local equating and kernel equating were used. The criterion used from the local equating method was bias, as defined in van der Linden and Wiberg (2010) and Wiberg and van der Linden (2011). The criteria used from the KE approach were the kernel equating standard error of equating (SEE) as well as the percent of relative error (PRE) as defined in von Davier et al. (2004a). Bias was examined

by comparing the resulting transformations $\varphi(y)$ with the true transformations $\varphi^*(y; \theta)$ in Equation 2, i.e., we used Lord (1980) definition of bias in equating. For each of these methods, the bias function associated with transformation $\varphi(y)$ (i.e., Equation 8 and 12) was:

$$\text{bias}(\varphi(y); \theta) = \varphi(y) - \varphi^*(y; \theta), \theta \in \mathfrak{R}. \quad (14)$$

The only exception to the definition of Equation 14 was the local kernel equating conditional on an anchor-test score method. Because Y and A are random, its bias functions were evaluated taking expectations over Y and A given θ that is,

$$\text{bias}(\varphi^{LAK}(y); \theta) = E_{A|\theta}[\varphi^{LAK}(y) - \varphi^*(y; \theta)], \theta \in \mathfrak{R}. \quad (15)$$

In addition, we also calculated the standard error and root mean squared error based on the differences between $\varphi(y)$ and $\varphi^*(y; \theta)$, but since the results were entirely aligned with those for the bias, they were excluded from this paper.

SEE was defined in Equation 7. The idea behind the percent of relative error (PRE) index is that the distribution of equated scores of the new Test Form Y should match the distribution of the old Test Form X to which the equating was intended. The distribution of the equated scores was continuous and the distribution of X was discrete. Hence, the PRE compared several moments of the two distributions (up to 10 moments). We let the p^{th} moment of the distribution of scores X and $\varphi_Y(x)$ be denoted as:

$$\mu_p(X) = \sum_k (x_k)^{p_{s_k}} \text{ and } \mu_p(\phi_X(Y)) = \sum_j (\phi_X(y_j))^{p_{r_j}}, \quad (16)$$

and the PRE defined as:

$$PRE(p) = 100 \frac{\mu_p(\phi_X(Y)) - \mu_p(X)}{\mu_p(X)} \quad p \in \mathfrak{R} \quad (17)$$

as in von Davier et al. (2004a). We examined the SEE and PRE for all kernel methods, i.e., both the ordinary and local kernel methods.

Method

Two test forms and an external anchor set, i.e., X, Y, and A, were randomly sampled and assembled from an item pool from an existing real-world testing program. All three test forms had a length of $m = 40$, in accordance with previous empirical studies (van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011).

The design for the first proposed local kernel equating method, labeled calibration design, had a population of $N = 82,000$ test takers, with 2,000 test takers at each of the ability levels on a scale range of $\theta = -2, -1.9, \dots, 2.0$. This distribution was chosen to ensure that we had enough test takers at each θ , and especially that we had enough test takers at the lowest and highest total score levels.

For the other two local kernel equating methods (appropriate only for the NEAT design), we generated the total number-correct scores on each test form for a population of $N = 50,000$ test takers. These test takers were sampled from a $N(-0.5, 1)$ and $N(0.5, 1)$ for population P and Q , respectively. We used large simulated data sets mimicking real data, as this setup is known to give valuable information (Harris & Crouse, 1993). The choice of sample size and population distributions for the NEAT design allowed for comparisons with the results in previous local equating studies where the NEAT design has been used (van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011).

The conditional distributions $F_{X|\theta}$ and $F_{Y|\theta}$, required for the true equating transformation (Equation 2) were calculated at ability levels ranging from $\theta = -2, -1.9, \dots, 2.0$ using the Lord and Wingersky (1984) recursion algorithm. Loglinear modeling was used to presmooth the distributions in (local) kernel equating methods for the NEAT design. The models with the best fit to the data were used, following the criteria in von Davier et. al. (2004). The equating transformations themselves were obtained using linear interpolation (local equating methods) and Gaussian kernel continuization (kernel and local kernel equating methods). The bias in the equating transformations was evaluated at the true transformation of

the scale range of $\theta = -2, -1.9, \dots, 2.0$, closest to the test takers' sampled value of θ . Several of the bias functions reported in the figures below do not have data points over the whole range of x values, because we omitted the values of X that had a probability lower than 10^{-3} . In a real testing situation, only a few test takers would have any of the omitted values and their reported equated scores would not be trustworthy due to a lack of accuracy. Descriptive statistics for all test forms used in the empirical study are given in Table 1; the cases that are not manipulated are referred to as baseline cases, i.e., one for the NEAT design and one for the calibration design. More specifically the descriptive statistics include the means, standard deviations (SD), coefficient α of the test forms, as well as their correlation (r_{XI}) and item characteristics (item discrimination (a), item difficulty (b), and pseudo guessing (c)).

The shorter anchor test of $m = 20$ items was obtained by randomly sampling half of the items from the baseline anchor test which had 40 items. The more discriminating condition for Anchor Test A and Test Form Y was realized by multiplying their discrimination parameters by two. These were the same manipulations performed in van der Linden and Wiberg (2010). Note, some of the differences may appear large but the base line data was built from a real testing program and the manipulations of data was used to illustrate how the methods worked under different situations.

[Table 1 about here]

Results

We first present the results for the methods for the NEAT design and end with the results for the methods examined for the calibration design.

The bias functions for the two proposed methods for the NEAT design together with their kernel or local counterparts are shown in Figure 1. Note, the traditional chain and post-stratification kernel methods only have one equating transformation each. For the local and local kernel methods the results are displayed either for every other anchor test score a or

every other $\theta = -2, -1.9, \dots, 2.0$ and 2.0 . The omitted results follow in general the same pattern since all results are ordered with a or θ value.

[Figure 1 about here]

All local methods gave less bias than their kernel methods counterparts and the least amount of bias was produced both by local kernel equating with ability estimated from an anchor score and local equating with ability estimated from the anchor items. This is in line with the results in van der Linden and Wiberg (2010) who examined local equating methods for the NEAT design. The local kernel methods gave in general smoother curves than their plain local counterparts. As hypothesized, by introducing log linear presmoothing, the precision of the equatings at the ends of the score scale, where the data were sparse, was increased. This means that for each ability, the local kernel methods ran in general over a larger range of the test score scale than the local methods ditto. If we reduce the number of anchor items by 50%, i.e., using 20 items instead of 40 items (Figure 2), we also reduced the accuracy of the anchor test scores as a proxy of the ability which we try to measure with tests X and Y . As expected, the bias of the local methods conditioning on anchor test scores increased when anchor test length was reduced, and its size was now comparable to the kernel equating methods. All local kernel equating methods were remarkable robust against changes in the number of anchor items. A reason for this might be that the kernel methods include presmoothing which is essentially a modeling part that makes the results more stable.

[Figure 2 about here]

From van der Linden and Wiberg (2010) and Wiberg and van der Linden (2011), it was evident that the largest impact on the equating would be produced by the changes in accuracy of the anchor test. This was exactly what happened when we changed the accuracy of the anchor test by using more discriminating items (Figure 3). All local kernel methods and plain local methods gave better results relative to the baseline cases (Figure 1): the local

kernel methods became more stable and their bias almost vanished, especially for local kernel equating conditioning on anchor scores. On the other hand, the chain and post-stratification kernel equating methods showed about the same amount of biases. These results are entirely in line with those of van der Linden and Wiberg (2010), who also found a remarkable decrease in bias for local equating methods with a NEAT design while traditional chain and post-stratification equating (i.e. not kernel versions) had the same amount of biases. Finally, if we would have changed the difficulty of the anchor test items instead of their discriminating power, the anchor test scores would essentially have kept their accuracy, and we would have got almost identical bias functions. Hence, this option was omitted.

[Figure 3 about here]

The percent relative errors were examined for all kernel methods and they showed only small differences with anchor test length or accuracy of the anchor test scores. In general, all methods had a small percent relative error for the first ten moments. Table 2 gives the percent relative error for the baseline case for the NEAT design; all other percent relative error tables are omitted due to their similar appearance.

[Table 2 about here]

In order to complete our comparisons of methods for the NEAT design, we also evaluated the kernel methods using their standard error of equating. The chain and post-stratification kernel equating standard errors were generally low, i.e., less than 0.2 in all conditions for all test scores; hence their graphs are omitted. The other methods had generally higher standard errors than chain and post-stratification kernel equating across all conditions. Since the differences between these methods were similar for each condition, only the standard errors for the baseline case for the NEAT design are shown in Figure 4 (top panel).

[Figure 4 about here]

Finally, Figure 5 illustrates our results for local and local kernel IRT observed-score equating as well as the IRT observed-score equating methods for the calibration design. Local kernel IRT observed-score equating gave small amount of bias although all methods behaved quite similarly, except for the fact that the kernel IRT observed score equating method yielded curves over the whole range of test scores. The last result is not surprising since the latter involves only one equating transformation for which all information is used. This method also involves a model-based approach and is therefore less dependent on the range of test scores than local (kernel) equating with estimated ability.

As Figure 5 illustrates, all methods were affected similarly by the simulated change in accuracy of the tests. The standard errors of equating for the IRT observed-score equating methods were generally low for both simulated conditions; hence, only the baseline case is shown (Figure 4, bottom). The standard errors for kernel IRT observed-score equating were lower than 0.2 for all test scores, except at the endpoints. These results were similar to those obtained for chain and post-stratification kernel equating with the NEAT design. The standard errors for the calibration design should be compared with those for the NEAT design (Figure 4, top). The former generally had smoother curves.

[Figure 5 about here]

The percent relative error was in general low and almost the same for all the IRT methods, ranging between $-.0767$ and $.1174$ for the local kernel IRT observed score equating and between $-.0042$ and $.3437$ for the kernel IRT observed score equating. It was also robust against changes in accuracy; hence their tables are omitted here. The fact that the percent relative error was generally low indicates that the equating tended to preserve the moments of the score distributions. These findings indicate that local kernel IRT observed score equating is a good alternative to regular kernel IRT observed score equating.

Concluding Remarks

The aim of this paper was to give an integrated view of local and kernel equating. We have illustrated our view by proposing three local kernel observed-score equating methods. Their evaluation suggests that local kernel methods are a good alternative when equating test scores. The methods were more stable than their plain local counterparts and less biased than their kernel counterparts. If we compare the results for the local kernel methods with those for their plain local counterparts in the NEAT design in van der Linden and Wiberg (2010), the equating functions for the former are smoother and more stable and in general produced equated scores over a larger test score ranges, a result probably due to the model-based presmoothing used in the kernel framework. Noteworthy, the percent relative error was remarkable stable across all conditions for all three methods. For more accurate anchor test scores, all local and local kernel methods gave less biased results than their kernel counterparts. This finding re-emphasizes that it is extremely important how the anchor test is constructed. The local kernel methods were somewhat less robust against a decrease in the length of the anchor test, which was of course due to the fact that shorter anchor tests tend to produce worse proxies for the ability measured by the X and Y tests.

The IRT observed score (local) kernel methods performed in general well with respect to all evaluation criteria. Although the joint fit of an IRT model to the X and Y test may appear as a stringent requirement, several large-scale assessment and testing programs that use multiple test forms (e.g., NAEP, TIMSS, PISA, and TOEFL iBT®) do show satisfactory fit.

The pessimism in van der Linden and Wiberg (2010) as to the traditional chain and post-stratification methods was confirmed in the current study. Both had generally higher bias than any of the other methods. This result is not too surprising since using just one transformation as a compromise between all members of an entire family of transformations is bound to yield more bias. At one point, however, the traditional kernel equating methods did

outperform their local kernel counterparts, namely for the standard errors of equating. This result is explained by the amount of data from which these standard errors are calculated: from all observations for the traditional method but only those for each level of ability for the local kernel methods. One local method, however, showed a slightly different pattern. The local kernel IRT observed-score equating method had generally low standard errors of equating.

From a practical point of view, we are optimistic about the proposed local kernel methods. Their integration of local equating with kernel equating was shown to have led to two major improvements: (i) better accuracy relative to the earlier local equating methods and (ii) much less bias relative to the earlier kernel methods. In addition, they appeared to be robust, had small percentages of relative errors, and produced equating transformations meaningful over a larger part of the test-score scale.

Summing up, the new local kernel IRT observed-score equating methods presented in this paper performed well on all evaluation criteria and seems ready for use in practice. The other two methods appeared to be dependent on the choice of anchor test. They performed well provided that the anchor test was reliable and produced accurate scores.

References

- Andersson, B. Branberg, K., & Wiberg, M. (2013a). kequate: The kernel method of test equating. R package version 1.2.1. <http://CRAN.R-project.org/package=kequate>
- Andersson, B. Branberg, K., & Wiberg, M. (2013b). Performing the kernel method of test equating using the package kequate. *Journal of Statistical Software*. In press.
- Harris, D. J. & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6(3), 195-240.

- Janssen, R., Magis, D., San Martin, E., & Del Pino, G. (2009, April). Local equating in the NEAT design. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: methods and practices. (2nd ed.)*. New York: Springer.
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29, 285-307.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8, 452-461.
- Sinharay, S. & Holland, P. W. (2008). The missing data assumptions of the nonequivalent groups with anchor test (NEAT) design and their implications for test equating. (ETS Research Rep. No. RR-09-16). Princeton, NJ: ETS.
- Stocking, M. L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- van der Linden, W. J. (2000). A test-theoretic approach to observed score equating. *Psychometrika*, 65, 437-456.
- van der Linden, W. J. (2006a). Equating error in observed-score equating. *Applied Psychological Measurement*, 30, 355-378.
- van der Linden, W. J. (2006b). Equating scores from adaptive to linear tests. *Applied Psychological Measurement*, 30, 493-508.
- van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking*. Chapter 13: pp. 201-223. New York: Springer.

- van der Linden, W. J. & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*, 34, 620-640.
- von Davier, A. A. (2010). Test equating observed-scores: the percentile rank, Gaussian kernel, and IRT observed-score equating methods. Workshop presented at International Meeting of the Psychometric Society, Athens, GA.
- 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. (2013). Observed-score equating: An overview. *Psychometrika*. DOI: 10.1007/S11336-013-9319-3.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The kernel method of test equating*. New York: Springer-Verlag.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, 41, 15-32.
- Wiberg, M. & van der Linden, W. J. (2011). Linear local observed-score equating. *Journal of Educational Measurement*, 48, 229-254.

Table 1

Descriptive Statistics for the Test Forms Used in the Empirical Study, with Standard

Deviations for Item Parameters Given within Parenthesis.

Design	Test form	Mean	SD	α	r_{XA}	a	b	c
Calibration	X	23.12	6.75	0.81	-	0.72 (0.15)	0.08 (0.80)	0.17 (0.10)
	Y	20.09	5.85	0.74	-	0.62 (0.19)	0.71 (0.86)	0.16 (0.09)
More Discriminating	Y	18.83	8.03	0.74	-	1.24 (0.38)	0.71 (0.86)	0.16 (0.09)
NEAT	X	20.60	5.80	0.73	-	0.72 (0.15)	0.08 (0.80)	0.17 (0.10)
	Y	22.10	5.32	0.68	-	0.62 (0.19)	0.71 (0.86)	0.16 (0.09)
$m = 40$	A (pop. P)	19.93	5.30	0.69	0.71	0.70 (0.22)	0.13 (1.13)	0.15 (0.10)
	A (pop. Q)	24.45	5.43	0.72	0.70	0.70 (0.22)	0.13 (1.13)	0.15 (0.10)
$m = 20$	A (pop. P)	9.97	2.98	0.50	0.61	0.68 (0.20)	0.03 (1.33)	0.14 (0.08)
	A (pop. Q)	12.16	3.02	0.54	0.61	0.68 (0.20)	0.03 (1.33)	0.14 (0.08)
Anchor Test is more discriminating	A (pop. P)	18.83	6.86	0.69	0.78	1.41 (0.44)	0.13 (1.13)	0.15 (0.10)
	A (pop. Q)	25.32	7.08	0.72	0.76	1.41 (0.44)	0.13 (1.13)	0.15 (0.10)

Table 2

Percent Relative Error for the Post-stratification (PK), Chain (CK) and the Local KE

Conditioning on Anchor Score (LAK) for Different a Scores in the NEAT Design.

Moments	PK	CK A to X	CK Y to A	LAK $a = 8$	LAK $a = 14$	LAK $a = 20$	LAK $a = 26$	LAK $a = 32$
1	.0003	.0009	-.0025	.0167	-.0019	.0000	.0001	.0006
2	.0015	-.0001	-.0050	-.0146	-.0049	-.0007	-.0009	.0016
3	-.0010	.0025	-.0071	-.0475	-.0040	.0015	.0012	.0031
4	-.0084	.0090	-.0043	-.0801	.0058	.0087	.0076	.0054
5	-.0204	.0183	.0082	-.1269	.0281	.0220	.0180	.0089
6	-.0363	.0294	.0338	-.2075	.0663	.0425	.0318	.0014
7	-.0551	.0417	.0746	-.3405	.1224	.0707	.0481	.0021
8	-.0760	.0551	.1317	-.5414	.1980	.1070	.0658	.0305
9	-.0980	.0698	.2054	-.8220	.2942	.1518	.0840	.0043
10	-.1206	.0867	.2945	-1.1904	.4114	.2054	.1017	.0059

Figure 1. Bias functions for the NEAT design when $m=40$ (baseline case).

Figure 2. Bias functions for the NEAT design when $m = 20$.

Figure 3. Bias functions for the NEAT design when $m = 40$ and a more discriminating A test.

Figure 4. Standard errors of equating for the baseline case of the three proposed local kernel methods, for the NEAT design on top and the calibration design below.

Figure 5. Bias functions for the calibration design, with the baseline case to the left and a more discriminating Y test to the right.