Alternative Linear Item Response Theory Observed-Score Equating Methods

Applied Psychological Measurement
2016, Vol. 40(3) 180–199
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615605089
apm.sagepub.com

Marie Wiberg¹

Abstract

Item response theory observed-score equating (IRTOSE) is widely used in many testing programs. The aim of this study was to empirically examine three alternative linear IRTOSE methods compared with the traditional IRTOSE method and to discuss these methods in light of previously suggested alternatives. This contribution is both conceptual, by exploring three alternative methods that fit into the current observed-score equating framework, and empirical by comparing the methods through simulations and with real data. The results show that the local linear (kernel) IRTOSE methods yield low bias and low values on loss measures. However, using only a linear IRTOSE method results in excessive bias and cannot be recommended because of the ease with which IRTOSE with full distributions can be performed. An example using real data showed considerable differences in the equated scores with the alternative methods as well as in comparison with the traditional IRTOSE method. Practical considerations are given in the concluding remarks.

Keywords

kernel equating, local equating, IRT observed-score equating, linear equating, observed-score equating

Test equating is used to ensure that test scores from different test forms are comparable and that the scores can be used interchangeably. A well-known observed-score equating method that has been used extensively over the years is item response theory observed-score equating (IRTOSE). In IRTOSE, an item response theory (IRT) model is used to produce estimated distributions of the observed number-correct scores on the test forms. These estimated distributions are then equated with equipercentile equating methods (Kolen & Brennan, 2004). The widespread usage of IRTOSE can be attributed to the extensive use of IRT to model items in many large-scale assessments around the world. Although IRTOSE is a well-working equating method, a need for refining the method as well as the opportunity to increase its usage by incorporating new methods into the current observed-score equating framework are seen (von Davier, 2011). In the last couple of decades, several local equating methods have been

¹Umeå University, Sweden

Corresponding Author:

Marie Wiberg, Department of Statistics, Umeå School of Business and Economics, Umeå University, SE-901 87 Umeå, Sweden.

Email: marie.wiberg@umu.se

suggested, including local IRTOSE (van der Linden, 2000). In local IRTOSE, a family of IRTOSE transformations is used instead of a single equating transformation, which means that there is one IRTOSE transformation for each ability of interest. In the equating literature, it is suggested to use linear equating instead of equipercentile equating if the only way the score distributions differ across test forms is in the first two moments. Wiberg and van der Linden (2011) suggested the possibility of local linear equating methods, which means that they defined a family of linear equating transformations. However, they did not explore the possibility of local linear IRTOSE methods. Within the kernel equating framework (von Davier, Holland, & Thayer, 2004), a local kernel IRTOSE method was recently suggested although a linear possibility was not explored (Wiberg, van der Linden, & von Davier, 2014).

The aim of this article was to explore three alternative linear IRTOSE methods, to discuss them in light of existing methods, and to compare the three alternative methods with the traditional IRTOSE method through an empirical evaluation using both a simulation study and real data from a college admission test. A special interest was to examine whether the suggested methods are robust with a smaller sample size at each ability level than has been used in previous local equating method studies. In this article, the focus is on the equivalent group (EG) design, but replicating the study with the non-equivalent group with anchor test (NEAT) design should be straightforward. There also already exist other local equating methods specifically defined for the NEAT design (cf. Wiberg et al., 2014). This article is important because it both enriches the observed-score equating framework by including alternative linear IRTOSE methods and gives an empirical evaluation of the proposed methods.

The three hypotheses are as follows:

Hypothesis 1: The local equating methods would yield the least bias because local methods do not average over all the abilities as do traditional equating methods, which have previously been noted (Wiberg & van der Linden, 2011)

Hypothesis 2: The linear equating methods would be stable over the entire range of test scores compared with equipercentile methods because unless the samples are huge, the estimates of the observed-score distributions tend to be inaccurate, especially at the tails. If the bias due to assuming linearity is minor (i.e., there are only small differences in higher moments between test score distributions), one might accept some bias to obtain better mean square errors than when using equipercentile equating.

Hypothesis 3: The kernel equating methods would be overall more stable than the non-kernel methods because this is the case with the local kernel methods studied in Wiberg et al. (2014).

To define an equating transformation, assume that we have two test forms X and Y with scores of X and Y and that the two different distributions of X and Y have strictly increasing cumulative distribution functions (cdfs) $F_X(x)$ and $F_Y(y)$. As noted above, IRTOSE uses an IRT model to create estimated distributions of observed number-correct scores on each test form X and Y (Kolen & Brennan, 2004; Lord, 1980). Modeling the data with an IRT model can be seen as one form of pre-smoothing, and IRTOSE methods can also be viewed as a supplement to IRT true-score equating methods. The strength of IRTOSE is that it can be used with any equating design provided that the two test forms jointly fit an IRT model. The observed-score equating framework (von Davier, 2011) that the IRTOSE methods are part of includes traditional equating methods (Kolen & Brennan, 2004), local equating methods (van der Linden, 2000; van der Linden, 2006, 2011; van der Linden & Wiberg, 2010), kernel equating methods (von Davier et al., 2004), and the recently developed jointly local kernel equating methods (Wiberg et al., 2014). Different parts of the observed-score framework are used when discussing existing methods in comparison with the three alternative IRTOSE methods presented here.

The three alternative IRTOSE methods are all linear methods, and two of them are also local equating methods and could thus be compared with the local linear methods presented in Wiberg and van der Linden (2011). Their proposed local linear methods were, however, especially designed for either the NEAT or the single group (SG) design, while in this article, the focus is on the EG design. In the local linear method that Wiberg and van der Linden (2011) proposed for the NEAT design, the anchor test had a central role as a proxy of ability because the conditional mean and variances over anchor scores were used to obtain a family of equating transformations. In their proposed method for the SG design, the test scores from the new test form were used as a proxy for the ability, thus the conditional means and variances were obtained over the new test scores. These two methods are different from the local linear methods proposed in this article because the observed test or anchor scores in our methods are not directly used to obtain conditional distributions. In our proposed methods, the conditional mean and variances are obtained over the test takers' ability. In Wiberg et al. (2014), the focus was on local kernel equating methods using the full distributions in both the NEAT and EG designs. The methods proposed here will, therefore, also be discussed in light of their proposed methods.

The structure of the article is as follows: The traditional equipercentile IRTOSE method is discussed in the next section, and this is followed by descriptions of the kernel and local equating methods. The "Linear Equating" section discusses IRTOSE in linear equating and gives definitions of the three alternative IRTOSE methods. The "Empirical Study" section describes a simulation study and a real data example, and the results of these are given in the following section. The final section contains some concluding remarks.

IRT Observed-Score Equating

The most well-known IRTOSE method is the traditional equipercentile IRTOSE method. Throughout this article, the three-parameter logistic (3PL) IRT model is used. Test taker l is assumed to have ability θ_l , and the probability to answer item i correctly is defined as

$$P(U = 1|\theta_l) = c_i + (1 - c_i) \frac{\exp(a_i(\theta_l - b_i))}{1 + \exp(a_i(\theta_l - b_i))}, \quad (1)$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the pseudo guessing parameter (i.e., the item characteristics). It is possible to use the compound binomial distribution to generate distributions of observed number-correct scores on test forms X and Y given the test taker's ability (Lord & Wingersky, 1984). The observed-score distributions are then cumulated over a population of test takers to obtain the number-correct observed-score distributions for both of the two test forms. If the ability distribution has a finite number of abilities for $l = 1, \dots, N$ test takers, the density functions with

$$\hat{f}_X(x) = N^{-1} \sum_{l=1}^N f(x|\hat{\theta}_l) \text{ and } \hat{f}_Y(y) = N^{-1} \sum_{l=1}^N f(y|\hat{\theta}_l),$$

can be estimated, where $\hat{\theta}_l$ are the ability estimates. The two estimated marginal distributions of test forms X and Y obtained from these calculations ($\hat{F}_X$ and $\hat{F}_Y$) are then equated with equipercentile equating (Kolen & Brennan, 2004; Lord, 1980; van der Linden, 2011). The following notation is used for traditional IRTOSE:

$$IRTO = \varphi(y) = \hat{F}_X^{-1}(\hat{F}_Y(y)). \quad (2)$$

Kernel Equating

Kernel equating (von Davier, 2013; von Davier et al., 2004) has the overall aim of finding an optimal version of the equating transformation between tests X and Y for a target population T . Let test forms X and Y be administered to different samples of test takers P and Q , respectively. If the two samples are either the same sample or from the same target population, we have $T = P = Q$. To construct kernel IRTOSE methods, the five steps of kernel equating that are briefly summarized here can be followed.

1. *Pre-smoothing*: The raw score distributions are smoothed by modeling the data, and the model with the best fit is selected and used.
2. *Estimation of the Score Probabilities*: From the estimated score distributions in Step 1, that is, $F_X(x_j) = Pr(X \leq x_j; T)$ and $F_Y(y_k) = Pr(Y \leq y_k; T)$, the score probabilities of X and Y for population T are obtained through $r_j = Pr\{X = x_j; T\}$ and $s_k = Pr\{Y = y_k; T\}$ for a test taker randomly selected from T scoring $X = x_j$ and $Y = y_k$, respectively. A design function is used that transforms the smoothed raw score distributions into marginal distributions for T .
3. *Continuization of Discrete Distributions*: The discrete test score cdfs are made continuous to ensure the existence of an equating transformation. Linear interpolation is traditionally used in equating for continuization, but in kernel equating, a kernel (e.g., logistic, Gaussian, or uniform) is used instead. A Gaussian kernel has typically been used, which for X (and similarly for Y) is

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

where $\Phi(z)$ denotes the standard normal distribution function, h_X is the bandwidth, and (von Davier et al., 2004)

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

The optimal bandwidth h_X can be found by minimizing the penalty function,

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

where $\hat{F}_T'(x, h_X) = \hat{f}_{h_X}(x)$, that is, the probability density function and κ is a constant within the range 0 to 1. $A_j = 1$ if

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

where w_{pen} is a constant (usually set to .25), otherwise $A_j = 0$ (Lee & von Davier, 2011). There are also other methods to obtain optimal bandwidths, although they give similar results (Häggström & Wiberg, 2014).

4. *Equating*: From the two continuized cdfs, the equating from test form Y to X is

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

In kernel IRTOSE (Andersson & Wiberg, 2015; von Davier, 2010; Wiberg et al., 2014), the estimated probabilities from an IRT model are used within the kernel equating framework, and using Equation 6 it is defined as

$$KIRTO = \varphi(y) = \hat{F}_X^{-1}(\hat{F}_Y(y)), \quad (7)$$

where $\hat{F}_X$ and $\hat{F}_Y$ are the smoothed cdfs.

5. *Calculating Standard Error of Equating (SEE)*: The asymptotic SEE for random sampling from the target population is defined as

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

Local Equating

Local equating (van der Linden, 2000; van der Linden, 2006, 2011; van der Linden & Wiberg, 2010) was developed from Lord's (1980) definition of equity, which in general means that distributions of the test scores and the equated scores are indistinguishable. To achieve equity, we cannot use a single equating transformation for an entire population as is used in the traditional equating methods because such a transformation will always be population dependent and biased. Instead, a family of equating transformations, where each ability level is one member of this family, needs to be defined. Let the test forms X and Y measure the same ability θ . If θ is known, the conditional distributions of X and Y to obtain the family of true equating transformations can be used,

$$\varphi^*(y; \theta) = F_{X|\theta}^{-1}(F_{Y|\theta}(y)), \theta \in \mathfrak{R}, \quad (8)$$

where each θ is a member of the equating family (van der Linden, 2000). Because the true ability is seldom known, it needs to be estimated or obtained from some kind of proxy, for example, the test scores in an EG design or the anchor test in a NEAT design.

The first defined local IRTOSE method was *local equating conditional on ability*, also named *local equating with estimated ability* (van der Linden, 2000), which is referred to here as local IRTOSE because it builds upon IRTOSE. Another previously defined local IRTOSE is the recently defined local kernel IRTOSE (Wiberg et al., 2014). Both of these previously defined local IRTOSE methods are equipercentile equating methods.

Linear Equating

The fundamental assumption in linear equating is that the only way the score distributions differ across forms is in the first two moments, that is, the means and variances. Let μ_Y and μ_X be the observed-score means of Y and X and let σ_Y and σ_X be their standard deviations. The linear equating transformation from Y to X is then defined as (Kolen & Brennan, 2004, section 2.2)

$$\varphi(y) = \mu_X + \frac{\sigma_X}{\sigma_Y}(y - \mu_Y). \quad (9)$$

This is only the general definition of linear equating, and for different equating designs, the means and standard deviations have to be derived from the observed test scores using different sets of assumptions. Linear equating can be viewed as an approximation of equipercentile equating, although it is also a special case of kernel equating if bandwidth parameters that approach infinity are used (von Davier et al., 2004, chapter 4). Linear equating is especially useful with small samples, and the accuracy of the results is most important near the mean (Kolen & Brennan, 2004, p. 293, Table 8.5).

Linear IRTOSE

The first of our alternative IRTOSE methods is *linear IRTOSE*, which is a linear extension of the IRTOSE method. Although IRTOSE has been used extensively over the years, we are not aware of any declared definition of a linear (traditional) IRTOSE method. Recently, Chen (2012) mentioned a linear IRTOSE in the kernel equating context in a NEAT design. However, it should be emphasized that Chen did not show or describe a linear IRTOSE definition in the traditional setting as it is described here. One obvious possibility to create a linear IRTOSE method is to derive the mean and standard deviations from the previously described distributions. One defines the linear IRTOSE equating transformation as

$$LIRTO = \varphi(\hat{y}) = \mu_{\hat{X}_\theta} + \frac{\sigma_{\hat{X}_\theta}}{\sigma_{\hat{Y}_\theta}} (y - \mu_{\hat{Y}_\theta}). \quad (10)$$

If $p_i(\theta)$ and $p_j(\theta)$ be the probabilities of answering an item correctly given θ , then

$$\mu_{\hat{X}_\theta} = E(\hat{X}_\theta) = \sum_{i=1}^{n_i} x_i \hat{p}_i(\theta) \text{ and } \mu_{\hat{Y}_\theta} = E(\hat{Y}_\theta) = \sum_{j=1}^{n_j} y_j \hat{p}_j(\theta), \quad (11)$$

and

$$\sigma_{\hat{X}_\theta}^2 = Var(\hat{X}_\theta) = \sum_{i=1}^{n_i} (x_i - \mu_{\hat{X}_\theta})^2 \hat{p}_i(\theta) \text{ and } \sigma_{\hat{Y}_\theta}^2 = Var(\hat{Y}_\theta) = \sum_{j=1}^{n_j} (y_j - \mu_{\hat{Y}_\theta})^2 \hat{p}_j(\theta). \quad (12)$$

An advantage with linear IRTOSE is its simplicity and that fewer test takers are needed with a linear method than when an equipercentile method is used. This method could be a good alternative when the items fit an IRT model. It is easy to compute, and if there are no real differences in the higher moments, it could easily replace the traditional IRTOSE. A disadvantage, shared with all linear methods, is that it might be over simplistic if the test forms differ in more than the first two moments because using the traditional IRTOSE method accounts better for differences between test forms.

Local Linear IRTOSE

The second alternative IRTOSE method is *local linear IRTOSE*, which is a linear extension of the combination of IRTOSE (Kolen & Brennan, 2004; van der Linden, 2011) and local equating conditional on ability (van der Linden, 2000). In this case, use θ as the index and derive a family of equating transformations from the conditional distributions of X and Y given θ , $F_{X|\theta}$, and $F_{Y|\theta}$. The family of equating is defined as

$$LLIRTO = \varphi(\hat{y}) = \hat{\mu}_{X|\theta} + \frac{\hat{\sigma}_{X|\theta}}{\hat{\sigma}_{Y|\theta}} (y - \hat{\mu}_{Y|\theta}) \quad \theta \in \mathfrak{R}, \quad (13)$$

where the means and variances of the observed-score distributions of X and Y are

$$\hat{\mu}_{X|\theta} = E(X|\theta) = \sum_{i=1}^n \hat{p}_i(\theta), \quad \hat{\mu}_{Y|\theta} = E(Y|\theta) = \sum_{j=1}^n \hat{p}_j(\theta),$$

$$\hat{\sigma}_{X|\theta}^2 = Var(X|\theta) = \sum_{i=1}^n \hat{p}_i(\theta) \hat{q}_i(\theta), \quad \text{and } \hat{\sigma}_{Y|\theta}^2 = Var(Y|\theta) = \sum_{j=1}^n \hat{p}_j(\theta) \hat{q}_j(\theta).$$

When the test items have been calibrated, these means and variances are known and so is the family of transformations. In an actual equating, θ from the new test form for the test taker is estimated and the member from Equation 13 at this estimate to equate the observed score Y to the score X is used.

An advantage with the local linear IRTOSE method is that if the items fit an IRT model and one wants to perform a local equating, it is an easily performed method that should yield accurate equating transformations as long as the test score distributions do not differ too much in the higher moments. A disadvantage is that this method is not expected to perform better than local equating based on the full conditional distributions of $F_{X|\theta}$ and $F_{Y|\theta}$. The computational advantages of using Equation 13 instead of a family of equipercetile transformations calculated from the full conditional distributions using the Lord and Wingersky (1984) algorithm are negligible.

Linear Kernel IRTOSE

A key feature in kernel equating is the bandwidths. By increasing the bandwidths h_x and h_y , a “linear” equating can be obtained. It is possible to define all kernel equating methods as consisting of a linear part and a remainder part (von Davier, 2011, p. 16; von Davier et al., 2004, p. 12):

$$\varphi(y) = Lin_Y(y) + R(y). \quad (14)$$

In general, increasing the bandwidths in kernel equating compared with applying the traditional linear equating in Equation 9 yields very small differences. If the bandwidth is larger than 10 times the standard deviation, one can refer to the kernel equating as linear (von Davier et al., 2004). There has also been a proposal for a unified approach to linear equating when using the NEAT design (von Davier & Kong, 2005). Note that if one has a NEAT design, Equation 9 cannot be applied in a straightforward manner. Instead, either linear chain or linear post stratification equating methods need to be used. In the kernel equating R package *kequate* (Andersson, Bränberg, & Wiberg, 2013), linear kernel equating can be obtained by using large bandwidths.

Local Linear Kernel IRTOSE

The third and last alternative IRTOSE method—*local linear kernel IRTOSE*—is a linear equating method that is a combined extension of IRTOSE (Kolen & Brennan, 2004), local equating conditional on ability (van der Linden, 2000), and kernel IRTOSE (von Davier, 2010). To obtain a local linear kernel IRTOSE, the local kernel IRTOSE method by using large bandwidths h_x and h_y is approximated. Local linear kernel IRTOSE is defined as

$$LLKIRTO = \varphi(y) = \hat{F}_{X_{h_X \rightarrow \infty} | \theta}^{-1} \left(\hat{F}_{Y_{h_Y \rightarrow \infty} | \theta}(y) \right), \quad (15)$$

where $h_X \rightarrow \infty$ and $h_Y \rightarrow \infty$ symbolize that large bandwidths are used. To perform the equating, θ for all test takers from the test forms can be estimated and a member of the family of transformations from Equation 15 is used to equate the observed score Y to the score on X . Because equivalent test takers are being used, that is, test takers with the same ability level, the identity function is the design function. The main difference between this method and linear kernel IRTOSE is that the former defines a family of transformations while the latter only gives one transformation.

The local linear kernel IRTOSE method for the sake of completeness is included. One might also argue that for linear equating, it is not meaningful to use pre-smoothing and continuization because the linear transformation is already continuous and the equating transformation is obtained from the mean and standard deviations of the score distributions. It is thus unnecessary to smooth these score distributions because they are never used in their entirety. The distributions are only used to estimate two moments (which might become worse because of bias introduced if smoothing is used). In von Davier et al. (2004), the problem of linear kernel equating is discussed and viewed as problematic because kernel equating is always equipercentile equating. Although there is a clear relationship between linear and equipercentile kernel equating (von Davier et al., 2004), the proposed method was viewed as a separate method here because it can be viewed as a linear equating method. Linear kernel equating is strictly theoretical and holds for a given kernel (e.g., Gaussian) when the bandwidths approach infinity. One can avoid over-smoothing by choosing large, but not too large, bandwidths so as not to introduce unnecessary bias. This being said, an advantage with the local linear kernel IRTOSE method is that fewer test takers are typically required for a linear method than when a full-distribution equating method is used.

Empirical Study

The goal of the empirical study was to evaluate the performance of the three alternative IRTOSE methods (linear IRTOSE, local linear IRTOSE, and local linear kernel IRTOSE) compared with the traditional IRTOSE method. Although for sake of completeness, other IRTOSE methods that are all related have been mentioned and the three new alternative IRTOSE methods in comparison with the traditional IRTOSE method have been evaluated because the other methods have been examined carefully before. The interested reader can refer to van der Linden (2000) and Wiberg et al. (2014) for details about previously suggested methods.

To achieve this goal, simulations where the conditions were known are used and the true equating transformations could be obtained. IRTOSE is closely connected to the definition of the true equating transformations in Equation 8, which is used, for example, in van der Linden and Wiberg (2010) and Wiberg and van der Linden (2011). Because of its close relationship with IRTOSE and the fact that it has worked well in previous studies, this definition was chosen here. The following three cases were examined: when nothing was manipulated, when a more difficult Y test was used, and when a more discriminating Y test was used. The latter two are known to be important factors when examining bias in equating methods (van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011; Wiberg et al., 2014). The empirical study was performed using MATLAB and R (R Development Core Team, 2014), and in particular the R package *kequate* (Andersson et al., 2013).

Evaluation Criteria

Each method was evaluated with respect to bias, percent relative error (PRE), and three loss indices. Lord's (1980) definition of bias was used to compare the obtained equating transformations (Equations 10, 13, and 15) with the true equating transformation in Equation 8:

$$\text{bias}(\varphi(y); \theta) = \hat{\varphi}(y) - \varphi^*(y; \theta), \quad \theta \in \mathfrak{R}. \quad (16)$$

Denote the p th moment of the distributions of X and $\varphi_X(y)$ by

$$\mu_p(X) = \sum_k (x_k)^p s_k \text{ and } \mu_p(\varphi_X(Y)) = \sum_j (\varphi_X(y_j))^p r_j,$$

and define the PRE as (von Davier et al., 2004)

$$\text{PRE}(p) = 100 \frac{\mu_p(\varphi_X(Y)) - \mu_p(X)}{\mu_p(X)}.$$

The three loss indices included the mean signed difference (MSD), the mean absolute difference (MAD), and the root mean squared difference (RMSD). Each equating method was compared with the true equating as defined in Equation 8. The indices were weighted by f_y —which is the observed frequency of scores on test form Y —and were defined as

$$\text{MSD} = \frac{\sum_{y=0}^n [\hat{\varphi}(y) - \varphi^*(y) f_y]}{N}, \quad (17)$$

$$\text{MAD} = \frac{\sum_{y=0}^n |\hat{\varphi}(y) - \varphi^*(y)| f_y}{N}, \text{ and} \quad (18)$$

$$\text{RMSD} = \sqrt{\frac{\sum_{y=0}^n [\hat{\varphi}(y) - \varphi^*(y)]^2 f_y}{N}}, \quad (19)$$

where N is the number of test takers given test form Y (Han, Kolen, & Pohlmann, 1997).

Method

Test forms X and Y were constructed by random sampling and assembled from an item pool of a large testing program. Throughout this study, the item parameters were assumed to be known with no estimation errors. The actual test is longer, but in this study, both test forms were chosen to have a length of $m = 40$ items. Although a different test was used, the chosen test length and basic settings in this study were in general the same or similar to previous empirical studies in the local equating field to enhance comparability (van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011; Wiberg et al., 2014).

A special EG design labeled the *calibration* design (cf. Wiberg et al., 2014) of $N = 82,000$ test takers for each test was used in which 2,000 test takers were at each of the ability levels on a scale range of $\theta = -2.0, -1.9, \dots, 2.0$. This special design was chosen to ensure that enough test takers at the lowest and highest total score levels are available. Linear equating does not actually require such large sample sizes, but these were chosen because one of our interests was in the estimation

of bias in the equating and to use large sample sizes minimize the problem with accuracy of the methods. Also, the actual test used is sometimes administered to such large groups, and the size of the samples was also in line with previous empirical studies within this field (Wiberg & van der Linden, 2011; Wiberg et al., 2014). To examine the performance of the explored methods with a smaller sample of test takers at each ability than has previously been studied in local equating, an additional calibration design was used with $N = 20,500$ test takers for each test in which 500 test takers were at each of the ability levels on a scale range of $\theta = -2.0, -1.9, \dots, 2.0$.

For the true equating transformations as well as for the different IRTOSE methods, the conditional distribution functions of the observed number-correct scores X and Y given different values of θ , $F_{X|\theta}$, and $F_{Y|\theta}$ are essential. The Lord and Wingersky (1984) recursion algorithm was used to obtain $F_{X|\theta}$ and $F_{Y|\theta}$ evaluated at each ability level $\theta = -2.0, -1.9, \dots, 2.0$. For the true equating, the known true conditional distributions could be obtained using Equation 8 by using known ability and known item parameters, meaning there were no estimation errors in the person and ability parameters. In the simulations, the test takers' item scores on X and Y were simulated using the known item and ability parameters with the 3PL IRT model in Equation 1. The model choice was motivated from the fact that the 3PL IRT model typically fits the items in the used test and the 3PL IRT model is one of the most commonly used in the testing industry and because we wanted to facilitate comparison with previous local equating studies. The simulated item scores were summed together to obtain test scores on X and Y , and they were also used to obtain estimates of the test takers' ability using the expected a posteriori (EAP) estimator. The bias was calculated as defined in Equation 16, that is, by comparing the true equating transformation (using the test takers' known true ability) and the proposed equating transformation for the test takers' closest estimated EAP value of ability at the points $\theta = -2.0, -1.9, \dots, 2.0$. Some of the bias functions in the "Results" section do not have data points at each score value because values of X with probabilities lower than 10^{-3} were omitted. To handle the discreteness of the test scores, the transformations at these values were obtained through linear interpolation in the traditional equating methods (Kolen & Brennan, 2004, section 2.5) and through Gaussian kernel continuization in the kernel methods (von Davier et al., 2004). Descriptive statistics of the test forms, including the mean, standard deviation, and item characteristics, are given in Table 1. The robustness of the methods in the equating was evaluated by examining the impact of the difficulty of the Y test and the discrimination of the Y test in relation to the X test.

The case in which nothing was manipulated is referred to as the baseline case in the tables and figures. The manipulations of the test forms were as follows. The more difficult test form Y was obtained by adding 0.5 to each of the baseline difficulty parameters in line with Wiberg and van der Linden (2011). The more accurate test form Y was obtained by multiplying the baseline item discrimination parameters by 2.0, which was in line with van der Linden and Wiberg (2010). These variations were chosen to determine the effects of these parameters on the bias in the equating, and common cases can, therefore, be found within these limits. Although kernel equating generally includes a pre-smoothing step with log-linear modeling, this was not used in this study because using an IRT model means that the data have already been smoothed and the authors did not want to over-smooth the data.

Real Data Study

To examine the practical aspects of the three explored methods, an example with real data from the Swedish Scholastic Assessment Test (SweSAT) was included. SweSAT is a college admission test that is given twice a year. It is a multiple-choice test with 160 items with eight subtests divided into a verbal and a quantitative section, both of which contain 80 items and are equated separately. In this example, an EG design was assumed because the SweSAT has typically been

Table 1. Descriptive Statistics for the Test Forms Used in the Simulation Study ($N = 82,000$) and in the Real Data Study.

Test form	<i>M</i>	<i>SD</i>	<i>a</i>	<i>b</i>	<i>c</i>
Simulation data					
X	21.41	6.20	0.82 (0.24)	0.76 (0.80)	0.25 (0.12)
Y	21.24	6.62	0.91 (0.26)	0.78 (0.66)	0.26 (0.10)
Y (more difficult)	19.01	6.01	0.91 (0.26)	1.23 (0.66)	0.26 (0.10)
Y (more accurate)	20.15	8.59	1.82 (0.52)	0.78 (0.66)	0.26 (0.10)
Real data					
X	37.95	13.45	1.24 (0.57)	0.65 (0.99)	0.15 (0.12)
Y	37.00	12.09	1.19 (0.49)	0.74 (1.10)	0.17 (0.13)

Note. Standard errors are within parentheses. *a* = item discrimination; *b* = item difficulty; *c* = pseudo guessing.

Table 2. Ability Categories With Interval Breakpoints and the Number of Test Takers Within Each Ability Category for Tests X and Y in the Real Data Study.

Ability	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2	
Break point*	−∞	−1.75	−1.25	−0.75	−0.25	0.25	0.75	1.25	1.75	∞
X	2,179	1,848	2,768	3,748	4,316	4,196	3,108	2,112	725	
Y	2,254	1,895	2,780	3,702	4,156	4,138	3,166	2,126	783	

*Note, an interval break point (e.g. −1.75) to the right of an ability (e.g. −2) belongs to that ability category.

equated with an EG design using different groups in the past (Lyrén & Hambleton, 2011). Two random samples both containing 25,000 test takers who took the quantitative section in the fall of 2011 (labeled Test X) and in the spring 2012 (labeled Test Y) were used. For information about the two samples and the item characteristics, please refer to Table 1. The test takers' estimated EAP abilities in the two samples were approximately normally distributed with a mean ability of -0.09 in both samples, and only a few test takers had very low or very high abilities. To avoid sparseness at some ability levels and to be able to use all of the data in the later analyses, nine ability categories were constructed by dividing the ability scale into nine intervals: $\theta = -2.0, -1.5, \dots, 1.5, 2.0$. The break points for these intervals as well as the number of test takers who belonged to each ability category for the two test forms are given in Table 2. Because this is a college admission test, the test is high stakes and is typically of greater importance for midrange and higher ability students. Because real data were used, the true equated values are not known, and thus, some of the previous evaluation criteria cannot be used. In the following "Results" section, equated values and PRE for $\theta = 1.5$ as an example are reported and the equating transformations in a figure for $\theta = -1.5, 0, 1.5$ are showed, that is, in the ability interval where most test takers were found. In general, a test taker with ability lower than -1.5 is unlikely to get into any educational program, and the equated scores for test takers with higher ability than 1.5 had similar patterns as the displayed equating transformations.

Results

Simulation Study

Only results for $\theta = -2.0, -1.0, 0, 1.0, 2.0$ are shown in the figures to enhance readability, but the omitted results follow the same pattern as the displayed results. In Figures 1 to 3 and Table

1, $N = 82,000$ test takers as opposed to Figure 4 where $N = 20,500$ test takers. Figure 5, Table 2, and the lower parts of Tables 1 and 3 show the results from the real data study with $N = 25,000$ test takers. The biases for the three proposed methods and for the traditional IRTOSE method in the baseline case are shown in Figure 1. As expected, both the local linear and local linear kernel IRTOSE methods had low bias. The linear IRTOSE method, however, had larger bias for the lower test scores, while the other methods were almost free of bias. This result is probably due to the use of a single equating transformation as opposed to the local methods that use a family of equating transformations.

For a more difficult Y test, as shown in Figure 2, all methods were found to have an increase in bias. This is not surprising because a linear method only adjusts for the differences between the first two moments of distributions, and the rest remains as bias. This means that when the relative difficulty of Y increases, the violation of the linear assumption becomes stronger. Similar difficulty level in the test forms yields better fulfillment of the linearity assumption, which is in line with what Wiberg and van der Linden (2011) noted when they studied other local linear methods in comparison with traditional linear equating methods. The fact that both traditional IRTOSE and linear IRTOSE had an increase in bias is not surprising because they are both derived from a similar theoretical foundation. It should be noted that with a more difficult Y test, the local methods performed better than linear and traditional IRTOSE with respect to bias.

For a more discriminating Y test, as shown in Figure 3, the (linear) IRTOSE methods displayed curves over the entire test score range, and the scores far away from an examined ability level had large bias. The local methods did not have curves over the entire test score range, but the bias was small for the three lower test abilities. The generally larger bias for all methods for the two highest abilities might be due to the fact that the examined test is designed to measure ability in the middle of the test score range, and it might have had problems capturing higher abilities. The local linear kernel IRTOSE method appeared to be similarly affected by changes in discrimination as its full-distribution counterpart examined in Wiberg et al. (2014), although it exhibited smaller differences in the baseline case for low abilities.

MSD, MAD, and RMSD were used to further examine the three alternative IRTOSE methods. In Table 3, the three cases with $N = 82,000$ test takers are given with the different evaluation criteria summed over the displayed categories. The local methods clearly scored lower in all three cases on all three measurements. This is not surprising because these measures compared the true equating against each of the equatings, and the local method gave lower values as expected because it compares the same ability levels against each other instead of comparing one equating transformation against each ability level. The lower part of Table 3 shows the PRE, which was in general lower for all local methods in the simulations. This result is probably due to the fact that it is easier to preserve moments within an ability level than when the entire ability range is used.

Finally, the simulation study was extended with a calibration design using a reduced sample size of $N = 20,500$ test takers, which is shown in Figure 4. Overall, the bias using either the large sample or the reduced sample was very similar, and thus, the cases of Y being more difficult and Y being more discriminating were omitted. The reason for this similarity is probably due to the fact that the three proposed methods are linear and having either 500 or 2,000 test takers within the same ability yields similar results in the mean and standard deviation within that ability.

Real Data Study

The test takers' test scores on test X and Y were approximately normally distributed in the real data study. Only equated values for $\theta = -1.5, 0, 1.5$ are shown in Figure 5 to enhance

Table 3. MSD, MAD, and RMSD for the Three Cases With $N = 82,000$ Test Takers. PRE for the Baseline Case and the Real Data Study.

Method	Baseline				Y more difficult				Y more discriminating			
	MSD	MAD	PRE2	PRE3	PRE4	PRE5	MAD	RMSD	MSD	MAD	RMSD	PRE10
IRTO	-0.600	1.498		1.058		0.522	2.263	1.512	-4.927	9.161	6.155	
LIRTO	-0.601	1.806		1.118		2.468	3.750	2.506	-4.250	10.64	6.376	
LLIRTO	0.192	0.463		0.645		1.222	1.334	1.394	1.902	2.319	2.471	
LLKIRTO	0.020	0.350		0.284		-1.344	2.646	2.212	-2.977	4.362	3.765	
Simulation												
IRTO	0.24	0.47		0.67	0.85	1.02	1.18	1.32	1.45	1.57	1.68	
LIRTO	0.11	0.21		0.42	0.73	1.12	1.53	1.93	2.29	2.60	2.83	
LLIRTO1	0.10	0.10		0.09	0.08	0.05	0.01	-0.08	-0.21	-0.40	-0.68	
LLIRTO3	0.05	0.01		-0.02	-0.03	-0.05	-0.06	-0.07	-0.08	-0.09	-0.12	
LLIRTO5	-0.06	-0.07		-0.09	-0.10	-0.12	-0.15	-0.18	-0.22	-0.27	-0.32	
LLKIRTO1	-0.00	0.00		-0.00	-0.00	0.02	-0.06	-0.13	-0.25	-0.35	-0.70	
LLKIRTO3	0.00	0.00		-0.00	-0.02	-0.03	-0.06	-0.10	-0.15	-0.21	-0.29	
LLKIRTO5	0.00	-0.00		-0.00	-0.01	-0.01	-0.02	-0.04	-0.06	-0.10	-0.13	
Real data												
IRTO	0.14	0.22		0.28	0.33	0.39	0.44	0.49	0.56	0.62	0.68	
LIRTO	0.01	0.02		0.18	0.60	1.30	2.23	3.37	4.68	6.14	7.74	
LLIRTO:1.5	-1.01	-2.03		-3.07	-4.12	-5.18	-6.24	-7.31	-8.39	-9.46	-10.54	
LLKIRTO:1.5	0.00	0.00		0.00	0.01	0.01	0.03	0.04	0.07	0.10	0.13	

Note. MSD, MAD, and RMSD in this table are summed over the displayed categories. The PRE numbers 1 to 10 refer to the 10 moments of PRE. The numbers 1, 3, and 5 in the method names for the simulation refer to the abilities $\theta = -2, 0$, and 2, respectively, for example, LLIRTO1 refers to LLIRTO for $\theta = -2$. In the real data study, 1.5 in the method name means that PRE is displayed for $\theta = 1.5$ for the LLIRTO and LLKIRTO methods. MSD = mean signed difference; MAD = mean absolute difference; RMSD = root mean squared difference; PRE = percent relative error; IRTO = traditional IRT; LIRTO = linear IRT; LLIRTO = local linear IRT; LLKIRTO = local linear kernel IRT.

Figure 1. Bias functions for five ability levels for the baseline case, $N = 82,000$.

Note. IRTOSE = item response theory observed-score equating.

readability, but it should be emphasized that the omitted results follow the same pattern as the displayed results. It should be noted that there are different equated values for low and high-ability test takers and that the traditional IRTOSE has equated values in between the high and low equating transformations of the local methods.

The equated scores from the four equating methods are given in Table 4. Equated scores for the local methods are only given for one ability level because the other ability levels gave similar size differences between the equated values. By comparing equated scores at some specific test scores, say 30, 50, 70, and 80, a few interesting details are evident. At scores of 30, 70, and 80, IRTOSE yields the lowest equated score. At 30, local linear IRTOSE, followed by local linear kernel IRTOSE, yields the highest equated score. At 50, local linear kernel IRTOSE, followed by local linear IRTOSE, yields the highest equated scores, and linear IRTOSE yields the lowest equated scores. At 70, local linear kernel IRTOSE and linear IRTOSE yield the highest equated scores. Finally, at 80, linear IRTOSE yields the highest equated score. It is evident that IRTOSE typically yields lower equated scores than the other methods and that the two local methods yield similar equated scores over the test score range compared with the other two methods. The most important observation from Table 4 is that depending on which equating method is used, there are differences that matter in the equated values, especially in the upper score range. This means that depending on which equating method is used, the test takers might

Figure 2. Bias functions for Y more difficult for five ability levels, $N = 82,000$.

Note. IRTOSE = item response theory observed-score equating.

have either lower or higher chances of being accepted into the college education program of their choice. The PRE (given in the lower part of Table 3) was in general low for all methods, although it was somewhat larger for the local linear IRTOSE, especially in the higher moments.

Concluding Remarks

This study aimed to explore three alternative linear IRTOSE methods, and these were evaluated in comparison with the traditional IRTOSE method with respect to bias, MSD, MAD, RMSD, and PRE. The two local methods performed well in many instances as seen in the figures compared with the traditional IRTOSE and linear IRTOSE method, with only a few exceptions in the lower observed scores. The local methods also performed better than IRTOSE on the MSD, MAD, and RMSD measures, except when the Y test was more difficult. Furthermore, the PRE was lower for the local methods in the simulation study although it was somewhat higher for local linear IRTOSE followed by linear IRTOSE in the real data study. The linear IRTOSE method performed worse in general compared with the traditional IRTOSE method, with a few exceptions. Because it does not require much extra effort to perform a full-distribution IRTOSE compared with a linear IRTOSE, the use of linear IRTOSE cannot be recommended in general.

Figure 3. Bias functions for Y more discriminating for five ability levels, $N = 82,000$.

Note. IRTOSE = item response theory observed-score equating.

It is mainly included here for the theoretical purpose of filling in a gap in the observed-score framework.

The two explored local linear IRTOSE methods had low bias, low values on the loss measures, and low PRE, which is promising. Previous research in local equating has mainly used bias and root mean squared error, but the latter was not included here because it yielded essentially the same results as the bias measure. By comparing the results of this study to those of Wiberg and van der Linden (2011), the two proposed local linear methods are at least performing as well as those suggested in their article. Furthermore, when comparing the results of this study with the full-distribution method counterparts given in Wiberg et al. (2014), we note that for low abilities, the linear versions perform better than full-distribution methods in the baseline case, and similar to changes in discrimination. When data that are well modeled by IRT is available, the proposed methods are especially preferable to the previously suggested local linear methods. Because many large-scale assessments use IRT to model items, it is well motivated to use an IRTOSE method.

One can always argue what is the use of more equating methods, but that gaps in the theoretical framework should be filled, and this is one of the main contributions of this study. IRTOSE is a stable equating method, and being aware that they have alternatives within the local

Figure 4. Bias functions for the baseline case when reducing the sample size to $N = 20,500$.

Note. IRTOSE = item response theory observed-score equating.

Figure 5. Equated values for the real data example with linear IRTOSE and IRTOSE in both graphs.

Note. To the left, LLIRTO, and to the right, LLKIRTO for the three abilities $\theta = -1.5, 0, 1.5$. IRTOSE = item response theory observed-score equating; LLIRTO = local linear IRTOSE; LLKIRTO = local linear kernel IRTOSE.

Table 4. Equated Scores for the Four Compared Methods in the Real Data Example.

Test Scores	IRTO	LIRTO	LLIRTO $\theta = 1.5$	LLKIRTO $\theta = 1.5$
0	0.0	-3.2	3.7	1.2
10	8.8	7.7	13.4	11.5
20	18.6	18.6	23.2	21.8
30	29.3	29.5	32.9	32.1
40	40.8	40.4	42.7	42.4
50	52.0	51.3	52.4	52.7
60	62.2	62.3	62.2	62.9
65	67.0	67.7	67.0	68.1
70	71.7	73.2	71.9	73.2
71	72.6	74.3	72.9	74.2
72	73.5	75.4	73.9	75.3
73	74.4	76.4	74.8	76.3
74	75.4	77.5	75.8	77.3
75	76.3	78.6	76.8	78.4
76	77.2	79.7	77.8	79.4
77	78.2	80.8	78.7	80.4
78	79.0	81.9	79.7	81.4
79	79.7	83.0	80.7	82.5
80	80.4	84.1	81.7	83.5

Note. The equated scores are given for $\theta = 1.5$ for the LLIRTO and LLKIRTO methods. IRTO = traditional IRTOSE; LIRTO = linear IRTOSE; LLIRTO = local linear IRTOSE; LLKIRTO = local linear kernel IRTOSE.

equating framework is an extra strength. There are also situations in which it is more appropriate to use local linear methods as opposed to local methods, for example, when the first two moments are of particular interest and the test forms do not differ in the higher moments. The reduced sample size simulation example pointed to the direction that the proposed methods can be used without information loss when we have smaller sample sizes at each ability level, which is a strength compared with methods that use full distributions. An obvious question of interest is how small the samples can be for the proposed local equating methods to still work. A simple answer would be that it depends on the number of used categories. If the sample size is small, one has the possibility to categorize the abilities into only a few categories. To examine how small each sample size can be at each ability category, a robustness study with different sample sizes and different numbers of ability categories should be performed. One guess is that one would need sample sizes at each ability category of interest of similar size as has been shown to be needed in other small-sample equating situations.

From the real data example, it is clear that the explored methods can be used in practice. It is also evident that choice of equating method affects the equated values and might have an impact on the test takers' results. The local linear methods are relatively easy to use, and they do not require as large sample sizes as local methods using full distributions. To circumvent the potential problem of too few test takers at an ability level and to be able to use all the data in the analyses, the test takers were categorized by estimated ability. This choice was made from a practical point of view and should be elaborated upon more in the future. The local methods can also be of specific importance to assure that specific ability groups achieve a fair equating, as noted in the real data example. From the real data study, larger differences in PRE were observed and this could possibly give us a hint for how to choose between local linear IRTOSE and local linear kernel IRTOSE, especially because the latter had more stable PRE results.

A strength with the three alternative IRTOSE methods presented here is that they can be directly fit into the current observed-score equating framework (von Davier, 2013) because they follow the steps outlined in that framework. The local linear kernel IRTOSE was derived from the joint methods for local and kernel equating, although all three explored methods were derived from linear equating and IRTOSE.

Two future challenges follow from this study. Because standard errors of equating have not yet been analytically derived for all of the proposed methods, they were not used as an evaluation criterion in this study. Although the possibility of using bootstrap standard errors has been recognized, it is believed that in the future, it would be of greater interest to derive these analytically and to make a thorough comparison of the standard errors for all (eight) existing IRTOSE methods. Another future challenge includes examining these methods with models other than the 3PL IRT model because M. von Davier, Gonzalez, and von Davier (2013) had indicated that some IRT methods might be somewhat problematic to use in local equating.

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: The research was partly financed by the Swedish Research Council grant 2014-578.

References

- Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating using the package *kequate*. *Journal of Statistical Software*, 55, 1-25.
- Andersson, B., & Wiberg, M. (2015). *Item response theory observed-score kernel equating*. Manuscript submitted for publication.
- Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating under the generalized kernel equating framework. *Journal of Educational Measurement*, 49, 269-284.
- Häggström, J., & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement*, 51, 201-211.
- Han, T., Kolen, M. J., & Pohlmann, J. (1997). A comparison among true- and observed-score equating and traditional equipercentile equating. *Applied Measurement in Education*, 10, 105-121.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- 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* (pp. 159-173). New York, NY: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8, 452-461.
- Lyrén, P.-E., & Hambleton, R. K. (2011). Consequences of violated equating assumptions using the equivalent groups design. *International Journal of Testing*, 11, 308-323.
- R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for statistical Computing.
- van der Linden, W. J. (2000). A test-theoretic approach to observed score equating. *Psychometrika*, 65, 437-456.

- van der Linden, W. J. (2006). Equating error in observed-score equating. *Applied Psychological Measurement, 30*, 355-378.
- van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking* (pp. 201-223). New York, NY: 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, July). *Equating observed-scores: The percentile rank, Gaussian kernel, and IRT observed-score equating methods*. Paper presented at the International Meeting of Psychometric Society, University of Georgia, Athens.
- 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* (pp. 1-17). New York, NY: Springer.
- von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*, 605-623.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.
- von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the nonequivalent group design. *Journal of Educational and Behavioral Statistics, 30*, 313-342.
- von Davier, M., Gonzalez, J., & von Davier, A. A. (2013). Local Equating using the Rasch Model, the OPLM, and the 2PL IRT Model—or—What is it anyway if the model captures everything there is to know about the test takers? *Journal of Educational Measurement, 50*, 295-303.
- Wiberg, M., & van der Linden, W. J. (2011). Local linear observed-score equating. *Journal of Educational Measurement, 48*, 229-254.
- Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement, 51*, 57-74.