<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., Gonzalez, J. (2016)

Statistical Assessment of Estimated Transformations in Observed-Score Equating

Journal of educational measurement, 53(1): 106-125

<https://doi.org/10.1111/jedm.12103>

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-119075>

Running head: *Assessing equating transformations*

Statistical assessment of estimated transformations in observed score equating

Marie Wiberg¹ and Jorge González^{2,3}

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

²PhD, Assistant professor, Department of Statistics, Pontificia Universidad Católica de Chile, Chile

³Measurement Center MIDE UC, Chile

Acknowledgement

The research in this article by Marie Wiberg was funded by the Swedish Research Council grant 2014-578 and Jorge González acknowledges partial support of Grant FONDECYT 11110044.

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

Running head: *Assessing equating transformations*

Cite as: Wiberg, M. & Gonzalez, J. (2016). Statistical assessment of estimated transformations in observed-score equating, *Journal of Educational Measurement*, 53(1), 106-125. <https://doi.org/10.1111/jedm.12103>

Statistical assessment of estimated transformations in observed score equating

Abstract

Equating methods make use of an appropriate transformation function to map the scores of one test form into the scale of another, so that scores are comparable and can be used interchangeably. The equating literature shows that the ways to judge the success of an equating (i.e., the score transformation) might differ depending on the adopted framework. Rather than targeting different parts of the equating process and aiming to evaluate it from different aspects, in this paper the equating transformation is viewed as a standard statistical estimator and from that viewpoint discusses how it should be assessed in an equating framework. For the kernel equating framework, a numerical illustration shows the potentials of viewing the equating transformation as a statistical estimator as opposed to assess it using equating specific criteria. A discussion on how this approach can be used to compare other equating estimators from different frameworks is also included.

Keywords: equating transformation estimates, statistical evaluation, equating-specific evaluation criteria, simulation

Statistical assessment of estimated transformations in observed-score equating

Equating methods are statistical tools used to ascertain that scores from different test forms are comparable and can be used interchangeably. The comparability is obtained using an appropriate transformation function that maps the scores of one test form into the scale of the other. To set up the problem, let X and Y be random variables denoting the scores from test forms X and Y with cumulative distribution functions (cdf's) $F_X(x)$ and $F_Y(y)$, respectively. Suppose that scores on X are to be transformed to the Y scale. A general transformation function used for the comparison of any two samples or distribution of random variables can be defined as,

$$\varphi_Y(x) = F_Y^{-1}(F_X(x)), \quad (1)$$

(Wilk & Gnanadesikan, 1968). In the equating literature, this function is known as the equipercentile transformation (e.g., Braun & Holland, 1992), and it has been shown that all the equating transformations are particular cases of it. Different equating frameworks lead to parametric, semiparametric and nonparametric estimators of φ (González & von Davier, 2013). Examples of frameworks include traditional equating methods (e.g. Kolen & Brennan, 2014), observed-score kernel equating methods (von Davier, Holland and Thayer, 2004; von Davier, 2013), item response theory methods (Lord, 1980; Kolen & Brennan, 2014), local equating methods (van der Linden, 2011), or combinations of these as in Wiberg, van der Linden & von Davier (2014). Because of the large number of possibilities, we need statistical tools to assess which equating estimator should be used in different situations.

Although from this setup it might appear evident that the main object of inference is the equating transformation, an obtained equating has typically been evaluated with different criteria within a particular framework. For example, in kernel equating, two measures which have been used are the percent relative error, which essentially compares the moments in the observed and the equated score distributions (see e.g. von Davier et. al, 2004; Jiang, von Davier & Chen, 2012), and the standard error of equating, which is known as a measure of the uncertainty in the estimated equating transformation due to the use of random samples (von Davier et. al., 2004). Within local equating the bias, i.e. the difference between the obtained equated values and the true equated values (see e.g. van der Linden & Wiberg, 2010; Wiberg & van der Linden, 2011) has been used, whereas studies using traditional equating methods have reported the so-called difference that matters, which originally was defined as the differences between equated scores and scale scores that are larger than half of a reported score unit (Dorans & Feigenbaum, 1994). All these measures, which we call *equating-specific*

evaluation criteria, target different parts of the equating process and aim to evaluate it from different aspects. For a summary of traditional equating criteria and how they have been implemented, see Harris & Crouse (1993) and Kolen & Brennan (2014). Harris and Crouse (1993) pointed out that there is not a single criterion which is preferable but rather several can be used. Although their review dates far back this is still true today as noted in Kolen and Brennan (2014, p. 325).

The view point taken in this paper is that the equating transformation is the parameter of interest to conduct statistical inference, and that estimations of it need to be statistically assessed as any standard evaluation of an estimator of an unknown parameter in statistics. This paper differs from previous studies of equating criteria as it explicitly shows how it is possible to define a true equating transformation against which the estimated equating transformation can be evaluated.

The aim of this paper is to discuss the statistical evaluation of equating transformation parameter estimates within an observed-score equating framework. Thus, this paper is an important step towards viewing the equating transformation as a statistical estimator (in line with van der Linden, 2011; González & von Davier, 2013). Although we will give general directions on how to handle a statistical assessment of equating transformations in various observed-score equating frameworks, the emphasis will be on kernel equating. This choice is mainly based on the fact that in this case a "true" value of the equating transformation (i.e., the parameter of interest) can easily be defined.

The structure of the paper is as follows. We first briefly summarize the kernel equating framework including a description of equating specific evaluation criteria within this approach. In the next sections, the general strategy for the statistical evaluation of equating transformation parameter estimates is presented and then discussed when used within four different equating frameworks, including kernel equating. Measures of statistical evaluation criteria to evaluate test scores are then described, followed by a numerical illustration. The final section contains some concluding remarks.

The Kernel Equating framework

Let x_j and y_k denote the possible score values that the random variables X and Y can take with $j = 1, \dots, J$ and $k = 1, \dots, K$, respectively. For test takers randomly selected from the target population T scoring $X = x_j$ and $Y = y_k$, define the score probabilities

$r_j = Pr\{X = x_j; T\}$ and $s_k = Pr\{Y = y_k; T\}$, respectively. Kernel equating (von Davier et al.,

2004) is an observed score framework (von Davier, 2013) whose goal is to find an optimal equating transformation between the test scores X and Y for a target population T , and which comprises the following five steps:

- (1) *Presmoothing*. Statistical models are fitted to the raw data to estimate the score distributions.
- (2) *Estimation of the Score Probabilities*. Using the so-called *design function*, the estimated score distributions from step 1, are transformed into estimated vectors of score probabilities $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$, for tests X and Y on the target population T , where $\mathbf{r} = (r_1, \dots, r_J)^t$ and $\mathbf{s} = (s_1, \dots, s_K)^t$.
- (3) *Continuization of Discrete Distributions*. Because test scores are mostly discrete, the cdf's need to be made continuous to ensure the existence of an equipercentile-like equating function (see Equation 1). The discrete random variable X is continuized by adding an independent random variable V together with a positive constant h_X as follows;

$$X(h_X) = X + h_X V. \quad (2)$$

In kernel equating one usually uses the Gaussian kernel, to obtain continuous test score distributions but other choices are possible too (e.g., a logistic or a uniform kernel). If the Gaussian kernel is used the continuized cdf 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 $\Phi(z)$ denotes the standard normal cdf, h_X is a bandwidth parameter controlling the degree of smoothing, and

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

In order to obtain an optimal bandwidth h_X it is common to minimize a penalty function, although other possibilities exist (e.g., Häggström & Wiberg, 2014; Liang & von Davier, 2014). The idea is to find a compromise between choosing a continuous distribution that should be as close as possible to the discrete distribution (score probabilities r_j), and at the same time, as smooth as possible without losing any fundamental characteristics of the distributions. When using a penalty function the bandwidth h_X is found by minimizing

$$\text{PEN}(h_X) = \sum_j (\hat{r}_j - \hat{F}'_T(x, h_X))^2 + \kappa \sum_j A_j, \quad (4)$$

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

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

where w_{pen} is a constant, usually set to .25, otherwise $A_j = 0$ (Lee & von Davier, 2011).

The continuization step is similar for s and Y .

(4) *Equating*. The equating transformation from test form X to test form Y is then defined as

$$\hat{\phi}_Y(x) = \phi_Y(x; \hat{r}, \hat{s}) = F_{h_Y}^{-1}(F_{h_X}(x; \hat{s}); \hat{r}) = \hat{F}_{h_Y}^{-1}(\hat{F}_{h_X}(x)). \quad (5)$$

(5) *Calculating the Standard Error of Equating*. The (asymptotic) standard error of equating (SEE) due to random sampling from the target population T is defined as:

$$SEE_Y(x) = \hat{\sigma}_Y(x) = \sqrt{\text{Var}(\hat{\phi}_Y(x))}, \quad (6)$$

where $\hat{\phi}_Y(x)$ is defined in Equation 5 and the delta method is used to calculate the variance under the square root sign.

Equating-specific evaluation criteria within the Kernel equating framework

As mentioned before, the two equating-specific evaluation criteria percent relative error and SEE have been used repeatedly within the KE framework. In this section, they are briefly summarized together with DTM.

The percent relative error (PRE). Denote the p^{th} moment of the distribution of test scores Y and equated scores $\phi_Y(X)$ by

$$\mu_p(Y) = \sum_k (y_k)^p s_k \quad \text{and} \quad \mu_p(\phi_Y(X)) = \sum_j (\phi_Y(x_j))^p r_j,$$

respectively. The PRE is defined as

$$\text{PRE}(p) = 100 \frac{\mu_p(\phi_Y(X)) - \mu_p(Y)}{\mu_p(Y)}, \quad (7)$$

(von Davier et al., 2004). PRE is an evaluation of how well the distribution of equated values resembles the distribution of the original test score distribution by comparing the moments in the observed scores and the equated scores distributions.

The standard error of equating (SEE). For kernel equating the SEE as defined in Equation 6 can be more explicitly written in vector form as

$$SEE_Y(x) = \left\| \hat{\mathbf{J}}_{e_Y} \hat{\mathbf{J}}_{DF} \mathbf{C} \right\|, \quad (8)$$

where $\| \cdot \|$ denotes the Euclidean norm of the vector, $\mathbf{C}$ is a matrix such that $\Sigma_{\hat{\theta}, \hat{\theta}} = \mathbf{C}\mathbf{C}'$, and $\mathbf{J}_{e_y}$ and $\mathbf{J}_{DF}$ are Jacobian matrices related to the equating transformation and the Design Function, respectively.

Difference that matters (DTM). It is defined as any score difference which would make a change in score reporting when the test scores are rounded (Dorans & Feigenbaum, 1994). Originally this meant any differences between equated scores and scale scores that are larger than half of a reported score unit. DTM has usually been used as a criteria to decide between two transformations which differ in some respect (e.g., subpopulation vs total population transformation, equated scores obtained using one transformation vs those obtained by another, and preequating vs postequating results). In this paper DTM will be used in connection to the measure of differences of two equating estimators, which is defined formally in the numerical illustration section.

All the evaluations discussed so far share the common feature that they are especially developed to evaluate an equating although they handle different aspects of the performed equating. These evaluations are quite different from general statistical tools used to assess different estimators, as they do not rely on comparing a true value with an estimated value of a parameter of interest. It should however be noted that statistical measures such as root mean square error, mean absolute difference and mean signed difference has been used previously to evaluate an equating (Harris and Crouse, 1993;) although not in the way that will be implemented here. Bias and mean square error have also been used to evaluate equating functions with sometimes several replications, as noted in Kolen and Brennan (2014, p. 311). They do not however emphasize the comparison of an estimated equating with its true equating but rather compare a criterion equating against an estimated equating function. Previous studies have typically used large populations from which replicated samples has been drawn and used to calculate some of the above mentioned measures. This study differ from previous studies in that a probability model is used to generate the data and this is replicated several times as is usually done in statistical assessment studies.

Statistical evaluation of equating transformation parameter estimates

Because a point estimator is defined as any function of the sample, measures of the quality of an estimator have been developed in order to choose between different estimators. Most of these measures are based on the deviation from the estimated value of a parameter to the true value of it. We will describe three of these measures in the case when the equating

transformation is the unknown parameter of interest. At this point, it is important to emphasize that the test scores X and Y are random variables, and that the equating transformation, as any other estimator in statistics, is just a mathematical rule which become a random variable when observed data (i.e., realizations of random variables) are used to obtain an estimation. It is thus necessary to take expectations to evaluate the quality of this estimator.

Let $\hat{\varphi}(x)$ denote an estimator of $\varphi(x)$. The definitions of bias, mean square error (MSE) and root mean square error (RMSE) are well known and standard and are shown here explicitly for the case of equating transformations:

$$\text{Bias}(\varphi) = E_{f_\varphi} [\hat{\varphi} - \varphi] \quad (9)$$

$$\text{MSE}(\varphi) = E_{f_\varphi} [(\hat{\varphi} - \varphi)^2] \quad (10)$$

$$\text{RMSE}(\varphi) = \sqrt{E_{f_\varphi} [(\hat{\varphi} - \varphi)^2]}, \quad (11)$$

where the expectation is taken over the distribution of the equating function estimates f_φ . For similar definitions within the local equating framework, see Wiberg and van der Linden (2011). Because these expectations are not generally available in closed form, in practice randomly generated score data are used to calculate them using Monte Carlo simulation.

There are also loss indices which can be used as statistical evaluation criteria and have been used to evaluate an equating. Three commonly used ones are mean signed difference (MSD), mean absolute difference (MAD) and root mean square difference (RMSD) (Han, Kolen & Pohlmann, 1997; Harris and Crouse, 1993). These measures are particularly useful for comparing two different equating estimators within a framework (e.g., one estimator that uses a Gaussian kernel and the other using a logistic kernel within the kernel equating framework). They can also be modified to compare a true equating transformation against an estimated equating transformation, as will be illustrated later in this paper.

Assessing $\hat{\varphi}(x)$ within an observed-score equating framework

In order to use any of the statistical evaluation criteria described above, we need to define a true equating transformation. The true value of φ is obtained for the equating framework using a particular definition of equating and as mentioned before we used Braun and Holland (1982) definition as stated in Equation (1) in this paper. It should be noted that, depending on the adopted framework, the mathematical form of the transformation φ might change. As long as the interest is to assess the equating transformation *within* an observed-score

framework this should not yield too much difficultness as the mathematical form of the true estimator and the estimated one is the same. However, when assessing *between* frameworks, the analytic form of φ will be different for the frameworks being compared which make the assessment task more difficult.

Once the true transformation is defined, simulated data are used to evaluate different estimators. Simulations are typically used to approximate the sampling distribution of the estimator under a particular set of conditions. Examining different situations through simulations is very useful when assessing the behavior of different estimators before they are applied on real data. If instead of simulated, real data are used, the true value of the parameter is not known, but that problem is shared with the rest of the statistics world. The common way around this problem is to generate data which mimic the real data and examine the estimator in different situations and then apply the best estimator on the real data which resembles the examined situations.

In what follows the statistical assessment of $\hat{\varphi}(x)$ will be discussed for four observed-score equating frameworks. These are kernel equating, and three equating methods based on item response theory (IRT), which utilize the conditional distribution of scores given the ability (González, Wiberg & von Davier, in preparation). In particular, IRT true score equating, IRT observed score equating and local equating are discussed. In all these cases we will first define a true value of φ and then generate replicated score data to evaluate different estimators $\hat{\varphi}$. Although the assessment strategy is common for all methods, both the mathematical definition of the equating transformation $\varphi(x)$ and the data generating mechanism used to generate replicated data will differ.

Assessing $\hat{\varphi}(x)$ within observed-score kernel equating

In order to assess the estimated equating transformation $\hat{\varphi}(x)$ within kernel equating we need to define a true equating transformation $\varphi(x)$. The true transformation will depend on what we are interested to examine, for example different bandwidth selection methods, different kernels, or how the equating transformation behaves if we have a symmetric or a skewed test score distribution. Because the equating transformation might depend on more than one parameter, to define the true equating transformation one has the flexibility to decide which parameter should be fixed and which should be estimated. As mentioned before, for the assessment of the equating transformation we can either: i) create simulated score data for the two test versions from a known probability model, an approach which is used in the later

empirical example; or ii) we can fit a model to the real data and then obtain sample replicates from the large population of test data. If the latter is used, the first step is to select the best possible model as the true model, using a goodness of fit test, Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) and of course make sure that the model assumptions are fulfilled. From the true model of the real data the true parameters which are needed to obtain true equated scores can be obtained.

Once we have obtained the true equating transformation we need to estimate the equating transformations for the scenarios of interest. In order to make a fair comparison which doesn't favor any of the assessed estimators, we need to compare the estimated transformations against the true equating transformation as defined using either of competing models (e.g. models defined using different kernels). In the kernel equating framework we can define the estimated equating transformation as

$$\hat{\varphi}_Y(x) = \varphi_Y(x; F_x, F_y, h_x, h_y, \hat{r}, \hat{s}) = \hat{F}_{h_y}^{-1}(\hat{F}_{h_x}(x)), \quad (12)$$

where h_x , h_y , $\hat{r}$ and $\hat{s}$ are defined as before, and F_x and F_y are used to make explicit the dependence of $\hat{\varphi}$ on the kernel used. Suppose the interest is in assessing the use of different kernels. Because using different kernels have impact on the estimation of the bandwidths h_x and h_y (see Equation 4) these values need to be estimated in each replication for the different scenarios. The same is true for the score probabilities r and s . If one instead of comparing different kernels, wants to compare different bandwidth selection methods one could let the score probabilities be fixed and estimate the bandwidths using different bandwidth selection methods. These are just two examples of possible comparisons and the first will later be illustrated in this paper.

Assessing $\hat{\varphi}(x)$ within IRT true score equating

In IRT true score equating, for a given value of θ , a number right true score ξ on X is considered to be equivalent to a true score η on Y that is associated with the same θ . The relation between ability and true scores is determined by the following parametric equations

$$\xi = \sum_j p_j(\theta, \omega_{xj}) = T_X(\theta, \omega_{xj})$$

$$\eta = \sum_j p_j(\theta, \omega_{yj}) = T_Y(\theta, \omega_{yj})$$

where $p_j(\theta, \omega_j)$ is the response function of an IRT model (e.g., the one (1PL), two (2PL) or three (3PL) parameter logistic model), and ω_j is a parameter vector representing item characteristics (e.g., difficulty, discrimination, guessing). Once item parameters are estimated, a compact way to write the estimated transformation to equate true scores is

$$\hat{\varphi}(\xi) = \varphi(\xi, \hat{\omega}_x, \hat{\omega}_y) = T_Y(T_X^{-1}(\xi, \hat{\omega}_x); \hat{\omega}_y). \quad (13)$$

Using Equation 13 for different values of θ , a table with equated pairs (ξ, η) is obtained. To define a true equating transformation φ , the abilities θ and the item parameters ω are given known (fixed) values. Then, for a range of ξ values on the X scale the fixed values of θ and ω are used to obtain "true" η equated values. To assess the transformation in Equation 13 we generate data following an IRT model using the fixed values of ω_j and $p_j(\theta, \omega_j)$. Thus, estimators that differ, for instance, in the way the item parameters are estimated or the IRT model assumed, can be assessed using the replicated data. Although in practice in IRT true score equating, observed scores (X, Y) are used instead of the unobservable true scores (ξ, η) the described assessment procedure remains the same.

Assessing $\hat{\varphi}(x)$ within IRT observed score equating

In IRT observed score equating, an IRT model is used to produce an estimated distribution of observed number-correct scores on each test form, which are then used to equate scores using equipercentile methods. The equating transformation is defined as follows

$$\varphi(x) = F_Y^{-1}(F_X(x)), \quad (14)$$

where $F_a(t) = \int F(a|\theta)g(\theta)d\theta$, $a = x, y$ and $g(\theta)$ is the distribution of θ (Kolen & Brennan, 2014). In IRT observed score equating it is possible to evaluate for example; different ability distributions $g(\theta)$, different IRT models and different ways to obtain the conditional distributions of scores $F(X|\theta)$ and $F(Y|\theta)$. If the interest is in comparing equating transformations obtained from using either 2PL or 3PL IRT models one needs to examine two possible true equating transformations, provided that the data is suitable to be modelled with both these IRT models. The true equating transformation will thus consist of Equation 14 with either the 2PL or with the 3PL. Simulated data can be obtained using a similar approach as described for IRT true score equating.

Assessing $\hat{\varphi}(x)$ within Local Equating

A more recent equating framework is local equating (van der Linden, 2011; van der Linden & Wiberg, 2010; Wiberg, et al, 2014). Instead of using the marginal distributions of scores, this method utilizes the conditional distributions of scores given ability or any simple classification of it, leading to a family of equating transformations of the form

$$\varphi(x; \theta) = F_{Y|\theta}^{-1}(F_{X|\theta}(x)), \quad \theta \in \mathfrak{R}. \quad (15)$$

Assessing $\hat{\varphi}(x) = \varphi(x; \hat{\theta})$ within local equating will be similar to IRT observed score equating if we use the local equating methods which rely on the assumption that the data fits an IRT model. We can then proceed and obtain estimates of the equating transformations similarly as in IRT observed score equating. Although various evaluation measures are possible to use, the assessment has typically been based on bias measures (e.g. van der Linden & Wiberg, 2010; Wiberg, et al., 2014). In this framework, what can be evaluate is for example different estimation methods for θ , different IRT models and different ways to obtain the conditional distributions.

Measures of statistical assessment when equating test scores

For each assessment criteria, the true and estimated equated values are compared for each test score. Let x_i denote a specific test score, where $i = 0, \dots, n$. The simplest evaluation criteria is the bias which for an equated value $\varphi_Y(x_i)$ over 1000 replications, where each replicate is denoted by l , is defined as

$$\text{Bias}(\varphi_Y(x_i)) = \frac{1}{1000} \sum_{l=1}^{1000} (\hat{\varphi}_Y^{(l)}(x_i) - \varphi_Y(x_i)), \quad (16)$$

followed by the mean square error (MSE) defined as

$$\text{MSE}(\varphi_Y(x_i)) = \frac{1}{1000} \sum_{l=1}^{1000} (\hat{\varphi}_Y^{(l)}(x_i) - \varphi_Y(x_i))^2, \quad (17)$$

and the root mean square error (RMSE) defined as

$$\text{RMSE}(\varphi_Y(x_i)) = \sqrt{\frac{1}{1000} \sum_{l=1}^{1000} (\hat{\varphi}_Y^{(l)}(x_i) - \varphi_Y(x_i))^2}, \quad (18)$$

where $\hat{\varphi}_Y^{(l)}(x_i)$ is the estimated equated score for the l th replication. It is also possible to use modifications of the previously mentioned loss indices; MSD, MAD and RMSD that use the

full set of replications. For that aim, we adjusted the loss indices to be used with replications and defined them as the following average loss measures:

$$\text{AMSD}(\varphi_Y) = \frac{1}{1000} \sum_{l=1}^{1000} \text{MSD}_l, \quad (19)$$

$$\text{AMAD}(\varphi_Y) = \frac{1}{1000} \sum_{l=1}^{1000} \text{MAD}_l, \quad (20)$$

$$\text{ARMSD}(\varphi_Y) = \frac{1}{1000} \sum_{l=1}^{1000} \text{RMSD}_l, \quad (21)$$

where $\text{MSD} = \frac{\sum_{i=0}^n [\hat{\varphi}_Y(x_i) - \varphi_Y(x_i)]}{n}$, $\text{MAD} = \frac{\sum_{i=0}^n |\hat{\varphi}_Y(x_i) - \varphi_Y(x_i)|}{n}$,

$$\text{RMSD} = \sqrt{\frac{\sum_{i=0}^n [\hat{\varphi}_Y(x_i) - \varphi_Y(x_i)]^2}{n}},$$

and n is the total number of items. These definitions are in line with how Chen (2012) used these measures although no formal definitions were given in that study. We have also explored the possibility to redefine these loss measures in order to use the full range of score values, rather than averaging over them. In this case, we started from the versions given in Han et al. (1997) and extended them into average loss at each score point. The proposed loss measures are named Average Point Signed Difference (APSD) and Average Point Absolute Difference (APAD) and defined as follows

$$\text{APSD}(\varphi_Y(x_i)) = \frac{1}{1000} \sum_{l=1}^{1000} [\hat{\varphi}_Y^{(l)}(x_i) - \varphi_Y(x_i)], \quad (22)$$

and

$$\text{APAD}(\varphi_Y(x_i)) = \frac{1}{1000} \sum_{l=1}^{1000} |\hat{\varphi}_Y^{(l)}(x_i) - \varphi_Y(x_i)|. \quad (23)$$

Note that it is also possible to define an Average Point Root Square Difference (APRSD) where the absolute value function in APAD is exchanged with a square root and the difference raised to two. As this would be mathematically equivalent to APAD its formal definition is excluded here.

Numerical illustration

In order to illustrate the statistical assessment of equating transformations, simulated data were used with a large number of replications to evaluate kernel equating using the Gaussian and Logistic kernels. The definition of the cdf using a standard Logistic kernel is similar to that used for the Gaussian kernel in Equation 2, but replacing $\Phi(\cdot)$ by

$$K(v) = \frac{1}{1 + \exp(-v)},$$

where V has mean 0 and variance $\sigma_V^2 = \pi^2 / 3$ (Lee & von Davier, 2011). A comparison between these two kernels using equating specific criteria (i.e., SEE) and cumulants was presented in Lee and von Davier (2011) where data from chapter 7 in von Davier et al. (2004) are used with the equivalent group (EG) design. For the numerical illustration, the reported values of $\hat{r}$, $\hat{s}$, h_x , h_y in these studies were used as true parameter values to obtain true equated scores. The score data come from two 20 items test forms X and Y where $M = 1455$ and $N = 1453$ test takers took test form X and Y respectively. For the true parameter values of the Gaussian kernel the optimal bandwidths were $h_x = 0.622$ $h_y = 0.579$ and for the logistic kernel the optimal bandwidths were $h_x = 0.512$ $h_y = 0.446$. Given the true parameter values for r_j and s_k , which are given in Table 7.4 (von Davier, et al., 2004), 1000 instances of score frequencies $\mathbf{n} = (n_1, \dots, n_J)$ and $\mathbf{m} = (m_1, \dots, m_K)$ were generated from the multinomial distributions $\text{Mult}(\mathbf{n}, r_1, \dots, r_J)$ and $\text{Mult}(\mathbf{m}, s_1, \dots, s_K)$ with the same sample sizes as in the original data. For each replication we estimated h_x and h_y as well as the r_j and s_k values. The first step in kernel equating is presmoothing. For simplicity we used the same loglinear model that originated the "true" score probabilities parameters to presmooth the simulated data as was given in von Davier et al. (2004)

$$\log(r_j) = \alpha_r + \sum_{t=1}^{T_r} \beta_{rt}(x_j)^t \quad \text{and} \quad \log(s_k) = \alpha_s + \sum_{i=1}^{T_s} \beta_{si}(y_k)^i,$$

with $T_r = 2$ and $T_s = 3$. Other possibilities are to either automatically select the loglinear models (using e.g. AIC, BIC, chi square test or likelihood ratio test) or skip the presmoothing step and argue that the same mistake is made in all replications. This could also have been done here although we have chosen to include the presmoothing step for illustrating reasons as it is part of the five steps of kernel equating.

From the generated data we estimated equated values for both the Gaussian kernel and the Logistic kernel. The estimated equating values in both these scenarios were then compared with the true equated values from both the Gaussian and logistic kernel yielding in total four comparisons or scenarios. This way, it is possible to check the robustness of misspecified models, for example those that use in the estimation a kernel which is different from the one used to produce the true equated values. The four different scenarios were examined with all the previously described statistical evaluation criteria (bias, MSE, RMSE, AMSD, AMAD, ARMSD, APSD, APAD) given in Equations 16-23. Additionally, adjusted versions of the PRE and SEE as well as an overall criteria of equating differences (EDIFF) were used and are described in the next subsection. To perform the actually numerical illustration either of the R packages kequate (Andersson, Bränberg, & Wiberg, 2013) or SNSequate (González, 2014) could be used. In this case we used SNSequate.

Adjusting the equating specific evaluation criteria

To take advantage of the full set of replications, we adjusted the previously described equating criteria PRE and SEE. To adjust SEE, we followed the general suggestion described in Kolen & Brennan (2014) and established an equating criterion using a number of replications in order to provide estimated equating relationships. For each score value x_i the SEE is obtained calculating the standard deviation of 1000 equated scores as

$$SEE_Y(x_i) = \sqrt{\frac{1}{1000} \sum_{l=1}^{1000} [\hat{\phi}_Y^{(l)}(x_i) - \bar{\phi}_Y(x_i)]^2}, \quad (24)$$

where

$$\bar{\phi}_Y(x_i) = \frac{1}{1000} \sum_{l=1}^{1000} \hat{\phi}_Y^{(l)}(x_i).$$

This approach has recently been used by Liang & von Davier (2014). Similarly, to adjust the PRE it is possible to average over replications, i.e. examine PRE_a

$$PRE_a(p) = \frac{1}{1000} \sum_{l=1}^{1000} PRE^{(l)}(p), \quad (25)$$

where $PRE^{(l)}(p)$ is the value of PRE as defined in (7) for the l th replicate. Finally the DTM or modifications of it has been used in several equating studies (e.g. Yang, Bontya and Moses, 2011; Brossman, 2010). Here, DTM is used as a criteria to decide between differences of two equating estimators measured by the equating difference (EDIFF) defined as

$$EDIFF = \frac{1}{1000} \sum_{l=1}^{1000} |\hat{\phi}_{Y_1}^{(l)}(x_i) - \hat{\phi}_{Y_2}^{(l)}(x_i)|, \quad (26)$$

where $\hat{\phi}_{Y_1}$ and $\hat{\phi}_{Y_2}$ represent two different equating estimators (Liang & von Davier, 2014).

Results

The bias, MSE and RMSE for the different equated values for the Gaussian and Logistic kernels using either of them as truth are shown in Figures 1 and 2. The suffixes (G and L) in these figures indicate which kernel (Gaussian or Logistic) was used to obtain true equated values, whereas the words *gauss* and *logis* indicate the kernel used for estimations. From Figure 1 it is evident that the bias at the endpoint is larger for misspecified models and it also differ depending which kernel is used. It should also be noted that the size of the bias is slightly different depending on whether the Gaussian or the Logistic kernel are used as the true equating transformation. Looking at the lower extreme (until score 7) the bias for gauss.G appears to be smaller and for both correctly specified models (i.e. gauss.G and logis.L) they are convergent from the middle to the upper part of scores. As expected, bias (at the endpoints) is largest for the case when the model is misspecified (logis.G and gauss.L).

Figure 1: Bias for equated values (over 1000 replications). G and L indicate that data were generated under the Gaussian and Logistic model, respectively.

Figure 2. MSE to the left and RMSE to the right for equated values (over 1000 replications). G and L indicate that data were generated under the Gaussian and Logistic model, respectively.

The MSE was almost identical in the four scenarios, except for slightly differences at the extremes of the score scale. This is not surprising as both the bias and the SEE (Figure 5) only displayed small differences across the score scale. The RMSE also yielded the same pattern for all scenarios and was never larger than .30 for any of the scores. The results for the average loss measures are shown in Table 1.

Table 1. Average loss measures for the four examined scenarios.

	gauss.G	logis.G	gauss.L	logis.L
AMSD	-0.00031	0.00361	-0.00419	-0.00026
AMAD	0.15866	0.16074	0.16040	0.15923
ARMSD	0.18147	0.18296	0.18283	0.18135

The suffixes (G and L) indicate which kernel (Gaussian or Logistic) was used to obtain true equated values, whereas the words *gauss* and *logis* indicate the kernel used for estimations.

From Table 1 it is evident that values of AMSD in all scenarios were close to zero when comparing Gaussian and Logistic kernel, indicating a small average loss regardless of the used kernel. In all cases, the values for misspecified models were larger than for the correctly specified models. The AMAD values indicates that the gaussian model performs better as it outperforms the logistic model when the model is correctly specified and it is also more robust when the model is misspecified. The values of ARMSD are all very similar across simulation scenarios. The results for the average point loss measures are summarized in Figure 3. There is clear consistency of the curves when correctly specified models are

evaluated across the full range of scores, which is the opposite conclusion for misspecified models, where particularly at the extremes of the score scale none of the models appear to be robust.

Figure 3. APSD and APAD (over 1000 replications) for the four examined scenarios.). G and L indicate that data were generated under the Gaussian and Logistic model, respectively.

These results are in line with what Figure 4 shows for the EDIFF measure. Although none of the differences is larger than a DTM, it can be seen that the largest differences between the gaussian and logistic model occur at the extremes of the scores scale.

Figure 4. The equating difference over the different test scores (over 100 replications), when comparing the Gaussian kernel against the logistic Kernel.

The SEE in the left part of Figure 5 yielded very similar results regardless of the used kernel except for small differences at the extremes. Interestingly, the SEE plot displays essentially the same results as that shown in Lee and von Davier (2011) (Figure 10.3) even though they did not use replicates. Notable is that the SEE curves in Figure 5 are almost identical to the RMSE curve shown in Figure 2. This might be due to the fact that for the former, the average of 1000 replicates used as criterion was very close to the true equating value in the latter.

A plot of the average PRE values for the first 10 moments of the distributions is shown in the right part of Figure 5. It is evident that average PRE is lower if a Gaussian kernel instead of a Logistic kernel is used, at least from the third moment on. Note, as $\mu_p(Y)$ in Equation 7 is the same regardless of estimated values there are only two curves in Figure 5 as compared with Figures 1-3.

Figure 5. SEE (average over 1000 replications) using either a Logistic or a Gaussian kernel to the left and the average PRE (over 1000 replications) using either a Logistic or a Gaussian kernel to the right.

Concluding remarks

This paper was motivated from the insight that test scores are random variables and thus an equating transformation is nothing else but a statistical estimator. This implicate that the obtained equating transformation, and then the equated scores, should be treated as estimates of a statistical estimator. Thus, if we want to assess an equating transformation we need to proceed as a standard statistical inference problem. The proposed approach to assess equating transformations has interesting features. First, it offers a statistical perspective on equating

which allows the usage of general statistical tools as e.g. probability generating models and replications in the statistical sense. Second, it emphasize that it could be fairer to use several true equating models instead of just one. Third, it gives us the possibility to use statistical evaluation methods which are already developed. A special challenge is that depending on which equating transformations are compared one must be careful not to favor any of those under comparison when selecting the true equating transformation. The data generation may also affect a comparison.

As we wanted to emphasize the statistical perspective on equating assessment we used simulated data from a known probability model. This not only allowed us to assess different components of the equating transformations (e.g., different kernels), but also it would allow to examine different equating methods in different (simulated) situations e.g. which equating method is most suitable if the score distributions are symmetric or non-symmetric. When it is known which equating method is most suitable in a particular situation, one can apply that method on the real data. However, when there is not an explicit probability model to generate data, real data can be fit with a model which is then used to obtain sample replicates from the large population of test data. This approach has been used in some studies as mentioned in Kolen and Brennan (2014, p.313) and references therein.

In this paper we have illustrated the assessment of an equating transformation within the kernel equating framework and discussed how this approach can be implemented within either the IRT true score and IRT observed score or local equating frameworks. In all these cases, an explicit probability model (i.e., the multinomial and the underlying Bernoulli model for binary IRT data) was available. A more challenging part is to assess the equating transformation within equipercentile observed-score equating because one has to decide which probability model can be used to generate the data. If we use a multinomial distribution we already have discrete test scores while if we decide to use for example a normal distribution we need to first discretize the scores. In any case, it is unclear which probability model better underlies the situation.

So far we have considered assessing $\hat{\phi}(x)$ only *within* an observed-score equating framework. But there are also situations when it is interesting to compare equating transformations from different observed-score equating frameworks. The case of assessing *between* different equating frameworks is however more complicated and difficult as the true value of the equating parameter differ between frameworks. This problem is not shared with regular statistics where the parameter indexing a probability distribution is a unique abstract

element with no specific mathematical form. How to proceed with a fair comparison in this setting will depend on which frameworks the methods come from. If we use for example IRT observed score equating the data needs to fit an IRT model. But, does that automatically mean that it favors the IRT observed-score framework or the related local equating over e.g. kernel equating or equipercentile equating? In local equating one has previously used a family of equating as the true equating (Wiberg & van der Linden, 2011). However, most of the local equating method requires that the test follows an IRT model. If we use this family as the true equating transformation when we compare between different frameworks we will probably favor local equating and also favor all IRT equating methods because the assumption of the data following an IRT model.

If one wants to compare equipercentile equating with the kernel equating framework, is it fairer to generate data with a multinomial model which probably suits kernel equating better, or, should we instead use normally distributed data which possible favor equipercentile equating? The latter might imply that one skips the presmoothing and continuization step or is it fairer to discretize the data and use all five steps in kernel equating? If one choose the latter it essentially means to compare linear interpolation vs. Gaussian kernel continuization. A recent reflection related to this is given in Chen (2012) who declared that observed differences between equating methods can come both from the framework used for equating and the particular equating estimator. Yet another challenge is if one wants to compare many equating transformations from different frameworks there might be a number of possible candidates for the true equating transformation making the comparison quite complicated. In this paper the focus was on how to assess the equating transformation *within* an observed score equating framework. An area for future research is how to make fair comparisons *between* different equating frameworks for different situations such as for example symmetric or skewed distributions. Chen (2012) concluded that more research is needed to determine the best equating practice and that there is a need to find a good and practical criterion to identify the most appropriate equating method. This paper can be seen as one step towards that direction. However we still have to work with the two most important questions: how should one define the true $\varphi(x)$ and how should one generate data to avoid an unfair comparison of methods? These questions are ongoing challenges as new methods are proposed and new comparisons are of interest.

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.
- Braun, H. I. & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp.9-49). New York: Academics.
- Brossman, B. G. (2010). Observed score and true score equating procedures for multidimensional item response theory. Doctoral thesis, University of Iowa.
- 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(3), 269-284.
- Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.
- González, J. (2014). SNSequate: Standard and Nonstandard Statistical Models and Methods for Test Equating. *Journal of Statistical Software*, 59(7), 1-30.
- González, J. & von Davier, M. (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement*, 50(3), 315-320.
- 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.
- Harris, D. J. & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6, 195-240.
- Häggström, J. & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement*, 51(2), 201-211.
- Jiang, Y., von Davier, A. A., & Chen, H. (2012). Evaluating equating results: percent relative error for chained kernel equating. *Journal of Educational Measurement*, 49(1), 39-58.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking: methods and practices*. (3rd ed.). New York: 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*. Chapter 10: pp. 159-173. New York: Springer.

- Liang, & T. von Davier, A. A. (2014). Cross-validation: an alternative bandwidth-selection method in kernel equating, *Applied Psychological Measurement*, 38(4), 281-295.
- 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.
- R Core Development Team (2013). *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for statistical Computing. ISBN 3-900051-07-0.
- 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. (2013). Observed-score equating: An overview. *Psychometrika*. 78(4), 605-623.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.
- Yang, W-L., Bontya, A. M., and Moses, T. P. (2011). Repeater effects on score equating for a graduate admissions exam. *Research Report ETS RR-11-17*.
- Wiberg, M. & van der Linden, W. J. (2011). Linear local 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*. 1, 57-74.
- Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*. 55, 1-17.