Applied Psychological Measurement
2022, Vol. 0(0) 1–18
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216221124087
journals.sagepub.com/home/apm
SAGE

Evaluating Equating Transformations in IRT Observed-Score and Kernel Equating Methods

Waldir Leôncio[1,2], Marie Wiberg[3], and Michela Battauz[4]

Abstract

Test equating is a statistical procedure to ensure that scores from different test forms can be used interchangeably. There are several methodologies available to perform equating, some of which are based on the Classical Test Theory (CTT) framework and others are based on the Item Response Theory (IRT) framework. This article compares equating transformations originated from three different frameworks, namely IRT Observed-Score Equating (IRTOSE), Kernel Equating (KE), and IRT Kernel Equating (IRTKE). The comparisons were made under different data-generating scenarios, which include the development of a novel data-generation procedure that allows the simulation of test data without relying on IRT parameters while still providing control over some test score properties such as distribution skewness and item difficulty. Our results suggest that IRT methods tend to provide better results than KE even when the data are not generated from IRT processes. KE might be able to provide satisfactory results if a proper pre-smoothing solution can be found, while also being much faster than IRT methods. For daily applications, we recommend observing the sensibility of the results to the equating method, minding the importance of good model fit and meeting the assumptions of the framework.

Keywords

equating, item response theory, classical test theory, psychometrics, simulation, statistics

[1] Department of Statistical Sciences, University of Padua, Padua, Italy
[2] Centre for Educational Measurement, Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway
[3] Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden
[4] Department of Economics and Statistics, University of Udine, Udine, Italy

Corresponding Author:
Waldir Leôncio, Domus Medica, Sognsvannsveien 9, Oslo 0372, Norway.
Email: w.l.netto@medisin.uio.no

Introduction

Equating Methods, Frameworks and Transformations

In the context of standardized testing, a particular test is often administered in different forms that may differ in content and administration time and space. Moreover, there is often the desire to compare these different administrations. Equating is a procedure that allows one to do so by mapping test scores from one test form onto the scale of another. Let X and Y denote the respective scores from test forms X and Y, and X be administered after Y. Assuming one is interested in transforming the scores from the new test X to the scale of the scores of the old test Y, there is a general transformation function commonly referred to as the equipercentile transformation (Braun & Holland, 1982) which allows for this mutual conversion. The formula for the equipercentile transformation is as follows:

$$\varphi(x) = F_Y^{-1}[F_X(x)]. \tag{1}$$

It is reasonable to expect that differences in data collection procedures or in the assumptions behind how they were generated could call for different equating approaches. Hence, different equating methodologies have been developed over time to cover common scenarios in real-life situations.

Wiberg and González (2016) used a statistical approach to show how one can compare equating transformations within a particular framework. While not stated explicitly, their work characterizes an equating framework as a well-defined design structure that leads to a particular type—that is, parametric, semiparametric, or nonparametric—estimator of φ. Established examples of equating frameworks include IRT Observed-Score Equating (IRTOSE, Lord (1980)) and Kernel Equating (KE, Von Davier et al. (2004)). Both frameworks usually operate on observed scores, but this paper will explicit use "OS" as part of the acronym of the former but not the latter to match the preceeding literature.

Equating frameworks are usually made up of different customizable parts. For example, the KE framework allows one to choose between a range of pre-smoothing functions and kernel functions, while IRTOSE offers models with up to four parameters measuring different item characteristics such as their discrimination (i.e., how well they differentiate between high-skill and low-skill examinees), difficulty, pseudo-guessing chance (a lower asymptote for the probability of correctly scoring an item) and an upper asymptote to account for items where scoring the item may never be a sure event, no matter how skilled the examinee is.

This paper follows the implicit definitions used by Wiberg and González (2016) by referring to each of the possible combinations above as a "method." For example, KE with a Poisson pre-smoothing model is one method of the KE framework. KE without any pre-smoothing is another method of the KE framework. Likewise, 2PL-IRTOSE and 3PL-IRTOSE are two different methods of the IRTOSE framework. Table 1 displays this information in an alternative way.

After illustrating the comparison of equating methods within the Kernel Equating (KE) framework and discussing how the same could be done *within* IRT Observed-Score Equating (IRTOSE) and local equating, Wiberg and González (2016) listed the evaluation of equating transformations *between* different equating frameworks as a problem to be solved on a different study. This article concentrates on this problem by showing one way to evaluate equating transformations coming from different frameworks. In particular, we will focus on how to evaluate equating transformations from IRTOSE, KE, and IRT Observed-Score Kernel Equating

Table 1. Examples of equating frameworks and methods.

Frameworks	Methods
KE	No pre-smoothing Poisson pre-smoothing + chain equating Poisson pre-smoothing + post-stratification equating ⋮
IRTOSE	1-parameter logistic model 2-parameter logistic model 3-parameter logistic model ⋮

(IRTKE, Andersson and Wiberg (2017)). Both simulated and real data will be used to conduct this study.

Statistical Tools to Compare Equating Methods

The challenge of comparing equating transformations might date as far back as to when the second equating method was created. General statistical tools—also known in the equating literature as "summary indices" (Harris & Crouse, 1993)—such as bias and mean squared error can be used for that purpose, but there are also equating-specific measures that can be employed to help compare the performance of different equating methods.

An example of equating-specific measure is the percent relative error (PRE), which compares moments in the observed and equated score distributions (Von Davier et al., 2004). In traditional methods, the difference that matters (DTM)—the difference between equated scores and scale scores that are larger than half of a reported score unit (Dorans & Feigenbaum, 1994)—is often used to equate differences between different methods.

When compared to summary indices, equating-specific evaluation measures often allow one to target different parts of the process and thus aim to evaluate the equating based on different but specific aspects. On the other hand, summary indices are more flexible and, arguably, more robust to the implementation of new methodologies, particularly those targeting the comparison of equating methods between different frameworks. In the end, even though different evaluation criteria exist, there is no single criterion which is overall preferable (Harris & Crouse, 1993, p. 230). Since even small decisions made at each step of the equating process could unintentionally endorse one particular framework, a perfectly-fair benchmark criterion might even be unattainable. Nonetheless, this is an important goal to pursue, and even small strides in this direction should be welcome.

An interesting question is whether KE works better than IRTOSE if the data are not generated from an IRT model. This begets another question about how to simulate item responses if we do not want to assume a particular underlying IRT model, as that might affect the equating results. A possible approach to handling these problems, which is used in this article, is to generate test data using probability distributions. Our novel IRT-free approach should yield comparisons which are less biased against non-IRT frameworks.

The rest of this paper is structured as follows. In the next section, the equating methods used are described. Then, statistical evaluation criteria are presented, with the chosen equating specifications being described. The fourth section exhibits the characteristics of the real and the simulated data used to apply the equating methods under study. The results are given in the fifth section, and the last section contains some concluding remarks.

Equating Methodology

The NEAT Design

Expanding the notation based on Kolen and Brennan (2014) and introduced in *Introduction*, let the observed data for a test form be organized in a matrix with I rows and J columns, where I is the number of examinees and J is the number of items on that form. Within this matrix we find dichotomous answers to each one of those $I \times J$ combinations.

To perform observed-score equating, two components must be known: the data collection design and the equating method used (Von Davier et al., 2004). Let P and Q be two independent populations of examinees of which samples of size I_{P} and I_{Q} are taken. Let X and Y be two unique sets of test forms, respectively containing J_{X} and J_{Y} items each, and A be a common test form containing J_{A} items. Test form $\mathrm{X}^+ = \{\mathrm{X}, \mathrm{A}\}$ will be administered to the sample from population P, and test form $\mathrm{Y}^+ = \{\mathrm{Y}, \mathrm{A}\}$ will be administered to the sample from population Q. This data collecting design is called non-equivalent groups with anchor test (NEAT, Von Davier et al. (2004)).

What follows is a brief description of the three equating frameworks studied in this paper. To facilitate the methodological comparison between them, please refer to Figure 1, which shows a bird's eye view of the frameworks where all test forms to be equated are contained within the same boxes.

What is not depicted in Figure 1 is that each form may receive a slightly different treatment when the equating procedure is actually applied. In IRTOSE, for example, rescaling is performed only on one of the forms (say, X^+, which in this case is also referred to as the "new form"), to bring it to the scale of the other, old form (Y^+). Rescaling is a necessary part of the IRTOSE procedure because the test forms contain different items, with different parameter estimates, which cause test scores to not be directly comparable. In KE and IRTKE, minutiae of the workflow change

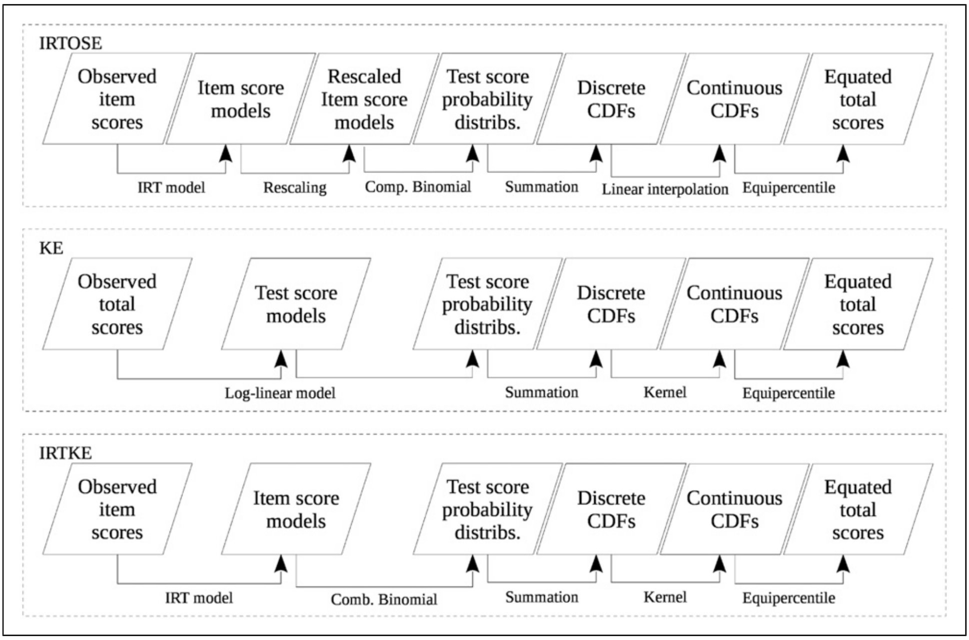


Figure 1. Simplified overview of three equating frameworks: IRTOSE, KE, and IRTKE.

depending on the equating method: under chain equating, for instance, form X will be equated to A before A is equated to Y.

Comparing these frameworks in general terms, what changes between them are the input data, the statistical model applied and the method used for making the score distribution continuous. Continuization, that is, the process of making the CDFs continuous, is one of the five steps of KE, dating back to Holland and Thayer (1987). In IRT equating, this step is not mandatory, but has the advantage of making score percentiles directly and bidirectionally comparable. If one were to use discrete scores, the equated form score distributions would differ (Kolen & Brennan, 2014, *Equating Choices for this Study*). In the context of this article, this step also allows one to better compare IRTOSE and KE.

After the continuous cumulative density function (CDF) of each test form is obtained, equation (1) can be used to equate the observed scores.

IRT Observed-Score Equating

The application of IRTOSE begins by fitting an IRT model to the test data. Assuming dichotomous items (i.e., items with only two possible answers, one of which being the correct one), the scores on test form X take values $X_{ij} = \{0, 1\}$, where i denotes the examinee and j denotes an item in form X. This comes with no detriment to the general definition of X as the random variable of the scores of X, and extension to test form Y and the respective scores Y_{ij} is straightforward. An IRT model calculates the probability of $X_{ij} = 1$ given the examinee's ability (θ_i) as well as some item parameters such as its discrimination (a_j) and difficulty (b_j). One of such models is the 2-parameter logistic IRT model (2PL), defined as

$$\Pr\left(X_{ij} = 1 \middle| a_j, b_j; \theta_i\right) = p_{ij} = \frac{1}{1 + \exp\left[-a_j\left(\theta_i - b_j\right)\right]}. \tag{2}$$

Let X_i be the number-correct score (or simply "score") of examinee i, i.e. $X_i = \sum_{j=1}^{J} X_{ij}$ so that $X_i \in \{0, \ldots, J\}$. A common way to calculate score probabilities is through a compound binomial model, here defined as

$$\Pr(X_i = x|\theta_i) = \sum_{\sum x_{ij} = x} \left[\prod_{j=1}^{J} p_{ij}^{x_{ij}} \left(1 - p_{ij}\right)^{1 - x_{ij}}\right]. \tag{3}$$

In practice, calculation is often performed through an iterative process described by Lord and Wingersky (1984), though other alternatives exist (González et al., 2016).

The test score probabilities in equation (3) are dependent on the abilities, so they must be marginalized to produce the probability distribution of the scores—$P(X_i = x)$—for a particular form, after assuming a distribution for θ_i. After continuization, equation (1) is applied on the resulting score distributions to obtain the equivalent scores (Kolen & Brennan, 2014, *Equating Choices for this Study and* section 6.6).

IRTOSE is a flexible equating method which can be used with any data equating design, provided that the two test forms jointly fit an IRT model (González & Wiberg, 2017).

Kernel Equating

KE is an equating framework comprised of five steps: pre-smoothing, estimation of score probabilities, continuization, equating, and calculating the standard error of equating (Von Davier et al., 2004, Ch. 3).

The goal of pre-smoothing is to fit a model to observed test scores so that design functions—as defined in Von Davier et al. (2004)—can be better used to estimate score probabilities. A useful family of pre-smoothing models are the log-linear models described in Holland and Thayer (2000): they are well-behaved, relatively easy to estimate and flexible enough to fit the types of score distributions that arise in practice (Von Davier et al., 2004).

Once the observed distributions have been pre-smoothed, the scores of the two forms need to be equated. This can be done through chain equating (CE) or post-stratification equating (PSE). Choosing between the two methods is still largely an open research topic, but differences between their results tend to be negligible when the populations P and Q have similar distributions to the anchor test form A or when A correlates highly with both X and Y (von Davier et al., 2004, section 11.8). In any case, the estimation of score probabilities is done by using a design function to transform the smoothed score distributions into marginal distributions for the populations.

In the third step, continuization of the discrete score distribution, KE uses a kernel—typically Gaussian, but also logistic, uniform or other—instead of linear interpolation. As with IRTOSE, once the cumulated score probability distributions have been made continuous, the two test forms can be equated by finding the scores located at the same percentile. Finally, accuracy measures such as the standard error of equating (SEE) can be calculated.

When compared to IRTOSE, KE offers the advantage of skipping the necessity to have the response matrix and to estimate parameters at the item level, not to mention the discussion about the most appropriate IRT model to use. Hence, KE has the potential to be less computationally intensive and more applicable than IRTOSE. On the other hand, it involves choosing (or not) a log-linear model to smooth the score probabilities, which arguably involves dealing with a greater number of model specifications than what IRT currently provides. Moreover, continuization requires selecting a kernel function as well as a smoothing parameter, which greatly increases the number of available combinations.

IRT Observed-Score Kernel Equating

IRTKE, described by Andersson and Wiberg (2017), uses score probabilities derived from an IRT model as input for kernel continuization. It can be seen as a compromise between the typical IRTOSE and KE procedures.

Since this method uses the IRT models from equations (2) and (3) on the pre-smoothing part of KE, it requires access to the item responses at the examinee level so that the model can be fit. When compared with CTT-based pre-smoothing models which only take into account the total test score of each individual, item-level models are more complex and consequently add computational overhead to the method.

This framework uses a continuous and differentiable kernel instead of linear interpolation for continuization of the score distribution, the latter of which is more common in IRTOSE methods. The reason for the change is that an everywhere-differentiable kernel allows the proper derivation of the asymptotic distributions (Andersson & Wiberg, 2017). Andersson and Wiberg also noticed that IRTKE works well for sample sizes as low as 1,000, as long as the 2PL model is used.

Equating Choices for this Study

Performing equating in one particular framework implies making several choices which result in one specific method within that framework. Picking only one method within each of the three frameworks under study is a decision that makes the number of comparisons manageable. One could argue that a fairer comparison between frameworks would require the evaluation of several methods for each framework, but our preliminary results have shown that the only case where changing the method gave wildly different final results was when a particular method generated blatantly unexpected equating results. This is usually a consequence of a model that clearly does not fit the data or simply fails to converge to a unique solution.

For IRTOSE and IRTKE, a 2PL model was fit to the item answers. When compared to alternative IRT models, the 2PL offers a good compromise between the simplicity of the 1PL and the flexibility of the 3PL. The flip side of simplicity is the potential failure to capture important characteristics of the items; on the other hand, flexible models with many parameters may have problems with convergence. Since IRT is performed on each test form separately, the estimated item parameters and abilities are on incomparable scales that do not reflect the relationship between the test forms they model (Kolen & Brennan, 2014, *On the Results Observed*). To handle this, the Stocking–Lord method (Stocking & Lord, 1983) was used to transform the item parameters.

To perform KE, we used a log-linear model fit through a Generalized Linear Model (GLM) for Poisson responses to model our data. Several log-linear models were considered, ranging from simple functions containing only the scores of the main and anchor tests as covariates to complex ones containing several powers of the partial scores (i.e., the scores of X, Y or A), the interactions between them, and dummy variables for low-frequency scores. The best model was then chosen by a stepwise method, which selects the model with the lowest Akaike Information Criterion (AIC). The equating between the scores of the two forms was done through CE, and continuization was achieved with a Gaussian kernel. The same choices were made for IRTKE.

The performance of a particular equating framework can be affected not only by the method, but also by how the data behave. Hence, we evaluated IRTOSE, KE, and IRTKE on three different data-generating scenarios: a Swedish college admissions test, a simulated test generated from IRT and another simulation with scores generated from a non-IRT procedure (Beta distribution).

All statistical procedures were performed in R (R Core Team, 2021), with ltm (Rizopoulos, 2006) being used to fit IRT models to the data and glm handling the log-linear models. IRTOSE was performed with equateIRT (Battauz, 2015); KE and IRTKE were done in kequate (Andersson et al., 2013).

Evaluating Equating Transformations

According to Wiberg and González (2016), the most common way to compare the performance of two equating methods *within* a particular framework is through equating-specific evaluation measures. Two popular examples are the DTM and the PRE, both of which could be adapted to compare equating transformations from different frameworks. However, such undertaking deserves a separate study and is thus not examined further in this paper, where more general alternatives will be employed.

In contrast to equating-specific measures, we could consider an equating transformation as a form of statistical estimator and calculate measures such as bias, standard error and root mean square error (RMSE). The advantage of this approach is the familiarity of such measures, and their application to a between-framework scenario seems straightforward. The employment of such measures to evaluate equating functions has previously been recorded in several published studies such as Wang and Brennan (2009); Lee et al. (2012); Lee and Brossman (2012); Wang and Kolen (2014); and Kim et al. (2019).

Considering the notation from Wiberg and González (2016), we respectively define the bias and RMSE for an equated value $\varphi(x)$ of score x over R replications as follows ($\widehat{\varphi}^{(r)}(x)$ is the estimated equated score for the r-th replication):

$$\text{bias}\left[\widehat{\varphi}(x)\right]=\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\varphi}^{(r)}(x)-\varphi(x)\right],\ \text{RMSE}\left[\widehat{\varphi}(x)\right]=\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\varphi}^{(r)}(x)-\varphi(x)\right]^{2}} \tag{4}$$

These equations make it clear that to calculate such measures we must have access to the true equating transformation $\varphi(x)$, which is not directly observable in real and most simulated data. There are, however, some ways to circumvent this limitation, one of which is to define one particular equating method as the true one and compare the others against it, something Wiberg and González (2016) did for different KE methods. This article uses two different approaches, one for real data and another for simulated data. These procedures are summarized in the following subsections, and the complete code can be obtained from the corresponding author upon request.

In addition to bias and RMSE, we provide definitions for the weighted average absolute bias (WAAB) and the weighted average root mean squared error (WARMSE), which are used in this paper to calculate averages of the statistics defined on equation (4) across the score range, weighted by score frequency accumulated across replications (i.e., $f_x=\sum_{r=1}^{R}f_{x_r}$):

$$\text{WAAB}=\frac{\sum_{x\in X}\left\{\left|\text{bias}\left[\widehat{\varphi}(x)\right]\right|\cdot f_x\right\}}{\sum_{x\in X}f_x},\ \text{WARMSE}=\frac{\sum_{x\in X}\left\{\text{RMSE}\left[\widehat{\varphi}(x)\right]\cdot f_x\right\}}{\sum_{x\in X}f_x} \tag{5}$$

Evaluating the Equating Results for the Real Data

To calculate evaluation measures for the real data, we used the same approach employed by Lord (1977, p. 132), which basically consists of composing test forms X and Y with the same items, while still having the computer handle them as being different. The procedure is summarized as follows:

1. The $I \times J$ matrix containing the examinees' answers to the J items on test form X was horizontally split into two matrices. The new matrices had half the number of examinees and the same number of items;
2. One of the resulting matrices was reassigned as test form Y;
3. Since X and Y are the same test form, equated scores should not change, i.e., $\varphi(x) = x$.

To reduce bias and improve the precision of our estimates, we have performed the steps above on 200 replications of the real data test.

Evaluating the Equating Results for the Simulated Data

The advantage of working with computer simulations is that they give the user control over the process that generates the data. The true equating scores are not explicitly defined by the data-generating process, but they can be obtained from it. Referring back to the section *Equating Methodology* and Figure 1, as long as the data-generating process allows the calculation of the expected CDFs of each test form, equipercentile equating can be performed.

For IRT-generated data, observations are generated by the true parameter values (a_j and b_j) and the true examinee ability (θ_i). Given these elements, the score probability CDFs were calculated as follows:

1. Equation (2) was applied to calculate $\Pr(X_j = 1|a_j, b_j; \theta)$;
2. These probabilities were used in the compound binomial distribution to calculate the probability distribution of the test scores given the ability, $\Pr(X = x|\theta)$;
3. By integrating θ out of $\Pr(X = x|\theta)$ multiplied by the density of θ (which is known), one obtains the unconditional score probabilities $\Pr(X = x)$, which can be cumulated to form the (discrete) CDF for one form.

Once the CDFs for forms X and Y are calculated, the real score equating transformation $\varphi(x)$ is found through equipercentile equating. Instead of converting the scores into a continuous scale before applying equation (1), this procedure applies equipercentile equating directly on the discrete CDFs. This prevents the challenge of choosing a fair continuization procedure (some procedures could favor one equating method over another). This is possible because the basic requirement for finding percentiles—i.e., reference values that split the data into two groups, one with lesser and one with greater numbers than the reference—is that the data can be ordered.

Using a continuous probability distribution such as the Beta distribution to generate test scores serves a double purpose: first, it provides a non-IRT method of generating test data, which is a welcome departure from how IRT procedures are often used in studies even when there is no intention of fitting IRT models or reason to believe the data meet its assumptions; second, it naturally bypasses the continuization issue discussed above, as the underlying test scores are already continuous and the calculation of $\varphi(x)$ is done directly, with no approximations.

Unlike what happens with the real data, the procedures above allow us to obtain all the elements necessary for calculating bias and RMSE as described on equation (4). In this study, R was set at 200, the point at which the final figures started offering stable results.

Real Data Application and Simulation Study

Real Data

As real data we used two administrations of the Swedish Scholastic Assessment Test (SweSAT): the autumn 2014 and the spring 2015 administrations. The SweSAT is a high-stakes, large-scale college admissions test given twice a year and is used for selection to higher education in Sweden. The test results are valid for 5 years and the examinees can retake the test as many times as they wish. Only their highest score is used when applying to colleges and universities. It is a multiple-choice, paper-and-pencil test consisting of a quantitative and a verbal section with 80 items each. The two sections are equated separately using anchor tests with additional 40 items each that are equated using KE with chained equipercentile equating. The common items are administered together with the unique ones. This study only uses the quantitative section so that a unidimensional model can be used. Sample sizes after this filter is applied are 2,826 and 2,783, respectively.

Simulated Data

To the extent of our knowledge, IRT models have often been used in the literature to generate test answers even when there was no intention of fitting an IRT model afterward. At the same time, the real world is not lacking examples of datasets which do not satisfy the necessary assumptions of IRT models. We believe differences in the data-generating process can affect the performance of equating, so the methods in this article were compared under scenarios containing data with different underlying data-generating procedures.

The simulated data are composed of two different tests: one with answers generated from randomly drawn IRT parameters and another with test scores generated from a Beta distribution. Each test contains 80 unique and 40 common items, mimicking the SweSAT.

The simulation begins by setting $I_P = I_Q = I = 1{,}000$ and generating a random vector of 1,000 examinee abilities for each one of the two forms. The ability distributions in those populations were set as $\theta_P \sim N(0, 1)$ and $\theta_Q \sim N(0.5, 1.2)$.

For the IRT data, we followed Andersson and Wiberg (2017) and generated a_j from a $U(0.5, 2)$ and b_j as $N(0, 1)$. With these item parameters and the true abilities θ, test answers were generated for test forms X^+ and Y^+ as Bernoulli trials with probabilities set by equation (2).

As an attempt to create a dataset that, at least in theory, should not fit an IRT model, a second group of tests scores was generated from a Beta distribution. Particularly, we set $X_i^* \sim Beta(2, 5)$, $Y_i^* \sim Beta(3, 6)$, and $A_i^* \sim Beta(2.5, 5.5)$. Correlation between scores of the main test (X or Y) and the respective anchor test A was set to 0.84, which is close to the one found on the studied SweSAT data (0.83).

In the literature, the Beta distribution has been used (together with the binomial or compound-binomial distributions as conditional distributions) to generate test data in conformance with the strong true score theory (Hanson & Brennan, 1990; Lord, 1965). Admissions tests such as the SweSAT are often positively skewed, so choosing a probability distribution that allows us to fine-tune the asymmetry of the score distribution is a sound idea, and the Beta fills the bill rather well. Since this study requires no need for conforming to strong true score theory, conditioning the Beta distribution on another one would just unnecessarily complicate the score generation process. All that we currently need is to retrieve scores that look realistic enough for using CTT-based methods such as KE (once again, we are trying to prevent the data from fulfilling the requirements for an IRT model).

The shape parameters for the Beta distributions were selected to give the test scores a fair amount of positive skewness, uncommon in IRT-generated data and yet often present in the real world (and, particularly, in the SweSAT data studied). Generating simulated asymmetric score distributions is possible using IRT models, but only indirectly (e.g. by having a large proportion of particularly easy or difficult items in relation to the examinee abilities). Using the Beta distribution, it is possible to directly influence this by simply changing the parameters: $Beta(\alpha, \beta)$ will be symmetric for $\alpha = \beta$ and asymmetric in the direction of the larger parameter.

Since the Beta always yields values between 0 and 1, the result was multiplied by the number of items (J) in that particular form to obtain a number between 0 and J, which is then rounded to make the observed test scores x_i, y_i, and a_i. This is how the random numbers generated by the Beta distributions are directly converted to test scores. It is a different approach from the IRT data, where item scores are generated first, and the test score is the sum of those item scores.

Having just test scores is enough to perform KE, but to fit an IRT model we must generate answers to each item. For the Beta data, this is achieved by generating, for each examinee, a vector containing x_i correct answers (1) and $J - x_i$ incorrect answers (0). Notice, however, that the order in which those 0s and 1s are generated will affect the item parameters, which will eventually be estimated by an IRT model. Simply generating, for each examinee, a sequence of 1s followed by a sequence of 0s may, once those vectors are stacked to form the answer matrix for all examinees, create many items which everyone answered correctly and many others which no one did, a situation in which IRT will be unable to estimate the item parameters. On the opposite end, uniformly scrambling those 0s and 1s will generate an answer matrix with no items standing out as particularly easy or difficult, which might be unrealistic. So naturally, a compromise had to be found.

In order to generate realistic test scores for the Beta data, each answer vector was permutated using probability weights that generate reasonable distributions of items according to difficulty.

Specifically, the weight vector $\boldsymbol{W}$ was calculated as $\boldsymbol{W} = (\boldsymbol{V} \circ \boldsymbol{V})/1000 + 1$. $\boldsymbol{V}$ is a vector of J equally spaced values ranging from $-J$ to J, where J is still the number of items in that particular test form. For instance, if $J = 5$, $\boldsymbol{V} = (-5, -2.5, 0, 2.5, 5)$ and $\boldsymbol{W} = (1.025, 1.00625, 1, 1.00625, 1.025)$. The "$\circ$" symbol denotes an entry-wise vector product. In practice, the permutated vector was obtained in R through the function sample(item_labels, prob = W), where item_labels is a vector containing the names of the items—for example, j1, j2, …, jJ—and W is the weight vector $\boldsymbol{W}$.

The function above ends up organizing the test data difficulty in a U shape, which is not a concern in our study because the order of the items should not affect the equating results. Moreover, when compared with other weight functions we have tested for this purpose (power functions, binomial and inverted normal distributions, slopes and V-shapes), the U-shaped weights ended up generating the most realistic sets of item parameters.

Results

Observed Real Data and Simulated Test Data

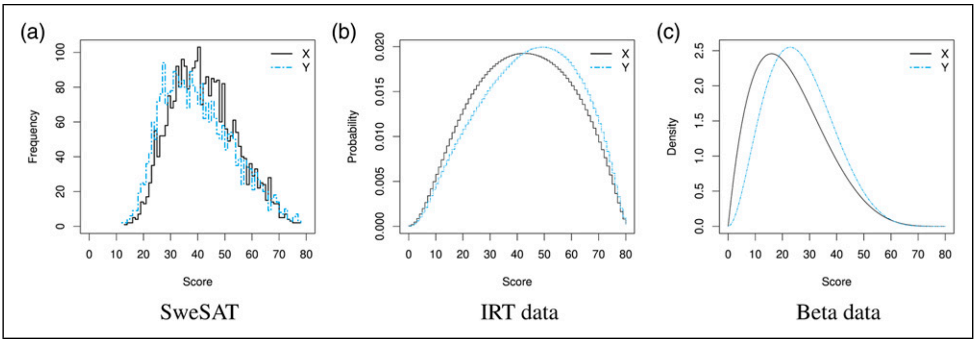
After developing and applying our novel method for generating test data using the Beta distribution, one of our first concerns was whether the generated values were credible, particularly when compared to the IRT-generated data and the real dataset under study. To examine this point, Table 2 and Figure 2, respectively, show descriptive statistics about the datasets and a graphical representation of the test score distributions.

When considering only the point estimates, the IRT-generated data show consistently higher average scores than the SweSAT, though that difference is arguably mitigated by the larger standard deviations of the former. The same can hardly be said about the Beta-generated data, which presents largely different average scores when compared to the SweSAT. It did, however, a better job replicating the correlation and skewness of the real data than the IRT-generated data did.

One other thing worth noting about Table 2 is how the results for the replicated SweSAT simulations have the same values for both forms. This happens as a result of the replication process, which involved randomly choosing one of the forms, splitting it in half and using each half as the two test forms to be equated to one another. As expected by the law of large numbers, given a sufficiently large number of replications the descriptive characteristics of the two test forms should converge to the same numbers.

Table 2. Descriptive statistics for the real and simulated data. Bars represent averages, $\widehat{\rho}$ and $\widehat{\gamma}$ respectively correspond to Pearson correlation coefficients and skewness, and numbers in parenthesis correspond to standard deviations. Results for the replicated and the simulated cases are averaged over 200 samples.

Statistic	SweSAT (original)	SweSAT (Replicated)	IRT data	Beta data
$\overline{X}^{+}$	58.4 (17.6)	57.3 (18.4)	60.7 (23.5)	35.4 (18.3)
$\overline{Y}^{+}$	56.5 (19.0)	57.3 (18.4)	71.9 (26.5)	39.2 (17.5)
$\overline{A}_X$	16.7 (6.4)	16.7 (6.5)	21.5 (8.1)	12.5 (6.2)
$\overline{A}_Y$	16.6 (6.6)	16.7 (6.5)	24.5 (9.1)	12.5 (6.2)
$\widehat{\rho}(X, A_X)$	0.82	0.83	0.92	0.84
$\widehat{\rho}(Y, A_Y)$	0.84	0.83	0.95	0.85
$\widehat{\gamma}(X, A_X)$	0.52	0.56	−0.02	0.54
$\widehat{\gamma}(Y, A_Y)$	0.60	0.57	−0.37	0.42



SweSAT IRT data Beta data

Figure 2. Distribution of observed test scores. For simulated data, values correspond to the theoretical score probabilities/densities.

The distribution of the observed scores for both the real and the simulated data can be seen in Figure 2. The score distributions of the anchor tests were omitted, as their similarity to the respective X and Y test forms makes them redundant.

The shapes of the score distributions in Figure 2 show slight positive asymmetry for the SweSAT and the Beta data, and slight negative asymmetry for the IRT data, impressions which are confirmed by the results posted on Table 2. For the simulated data, examinees taking test form Y have a slightly higher average ability than those who took X, which is different from what was observed in the real dataset. The differences, however, are well within one standard deviation and therefore deemed insignificant.

Equating Results

Information regarding the quality of the equating transformations can be seen in Figures 3 and 4, respectively presenting the bias and RMSE per score, data and method. These statistics were calculated according to equation (4). The figures show similar patterns for IRTOSE and IRTKE, which is expected since these methods only differ by their continuization algorithm as well as by how IRTOSE transforms the item parameter estimates of the new form, whereas IRTKE does not use them under CE. On the other hand, the behavior of KE shows more distinction, particularly on the SweSAT and the Beta-generated data.

The higher values observed in the [60, 80] score range for KE in Figure 4(c) are due to larger differences between the observed scores and their expected counterparts. The same applies to the [0, 20] and the [60, 80] score ranges on Figure 4(a). Unsurprisingly, these ranges match the score regions with very low frequency on Figure 2.

Table 3 supplements the graphical information provided by Figures 3 and 4. It summarizes the numerical results for the bias and RMSE, giving more insight into the comparisons. Some scores were omitted for brevity, but were still included in the averages. The absolute bias and RMSE are summarized by the weighted averages defined on equation (5). Using score frequency as weights gives higher importance to the most frequent parts of the score distribution.

Corroborating what was observed in the figures and using WAAB and WARMSE as the evaluation criteria, the table shows IRTOSE and IRTKE consistently outperforming KE. For the IRT data, KE offered less average bias than IRTOSE, but it did so at the cost of higher average RMSE. Nonetheless, it must be noted that the lack of a significance threshold on this kind of evaluation criterion paired with how the results from these tables often differ by less than one unit makes it difficult to point out a clear, universal winner. Moreover, when analyzing the table one

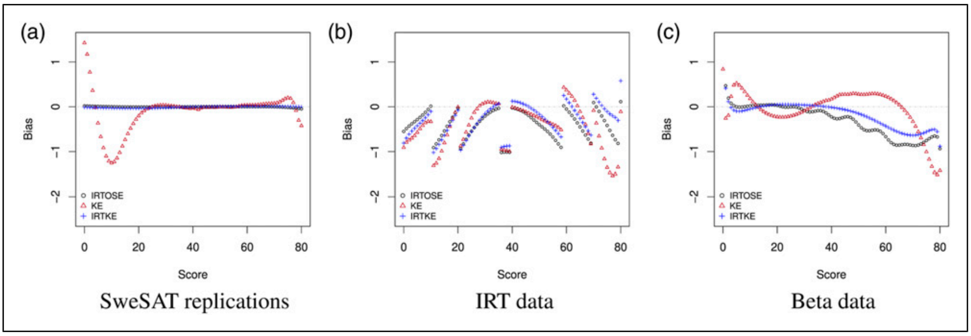


Figure 3. Bias per score.

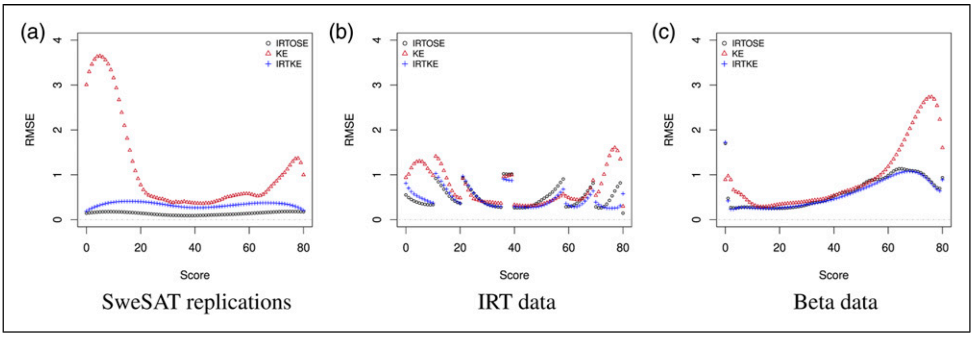


Figure 4. RMSE per score.

must consider how the different equating conditions and data-generation procedures could influence the results themselves, further hindering their direct comparison.

Discussion

Contributions of this Study

This paper expands on the work of Wiberg and González (2016)—focused on the comparison of equating transformations within the KE framework while also providing a general strategy for comparing transformations between frameworks—by applying a methodology to compare equating transformations from different frameworks, using IRTOSE, KE and the hybrid IRTKE as an example. Additionally, this study advances the work of Leôncio and Wiberg (2018) by implementing a framework for conducting computer simulations to evaluate the equating transformations under study. Additionally, this study proposes a new method to generate item responses without relying on IRT parameters so that comparisons between IRT and non-IRT equating methods can be done on fairer grounds.

Another approach for comparing equating transformations in the context of the NEAT design, as proposed by Sinharay and Holland (2010), is to order the equating transformations according to the difference between the Estimated Equating Function (EEF, or $\widehat{\varphi}(x)$) and the True Equating Function (TEF, or $\varphi(x)$). Our method differentiates itself from theirs by proposing an alternative way of calculating the TEF which is based on specific characteristics of the equating transformations studied instead of a general approach, potentially trading off scope for accuracy. As a

Table 3. Bias per score, Weighted Average Absolute Bias (WAAB) and Weighted Average Root Mean Squared Error (WARMSE) for the simulated and real data.

	SweSAT replications			IRT data			Beta data		
Score	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE	IRTOSE	KE	IRTKE
0	0.033	1.431	0.003	−0.551	−0.910	−0.810	1.698	0.839	1.723
10	0.009	−0.672	−0.002	0.012	−0.331	−0.102	0.019	0.182	−0.049
20	−0.000	−0.099	−0.010	−0.034	−0.005	−0.078	0.029	−0.226	0.046
30	−0.004	−0.002	−0.018	−0.208	0.091	−0.158	−0.084	−0.090	0.038
40	−0.008	−0.018	−0.030	−0.026	−0.009	0.126	−0.225	0.171	−0.013
50	−0.013	−0.021	−0.040	−0.400	−0.252	−0.167	−0.407	0.281	−0.151
60	−0.017	−0.095	−0.041	−0.051	0.372	0.172	−0.668	0.234	−0.432
70	−0.015	0.090	−0.024	0.095	−0.084	0.278	−0.864	−0.329	−0.636
80	−0.029	−0.434	−0.000	0.114	−0.110	0.579	−0.936	−1.422	−0.883
WAAB	0.008	0.034	0.028	0.403	0.372	0.326	0.084	0.181	0.050
WARMSE	0.111	0.476	0.315	0.500	0.563	0.450	0.349	0.431	0.337

matter of fact, our study uses not just one TEF but two different TEFs: one for the real data, using identity equating, and one for the simulated data, using the known real parameters to generate the TEF. One common drawback of ranking methods is that they could mask intricacies in the differences between the equating transformations, so once the EEF and TEF are calculated, we compared the bias and RMSE of each method directly.

The other innovation introduced in this paper attempts to solve a problem that should be addressed more often: using IRT models to generate test data even when there is no fundamental reason to use such restrictive models. Our proposal to generate data from a probability distribution shows satisfactory results which are easy to replicate and adapt. This overcomes a problem raised by Sinharay and Holland (2010), who stated that simulating data from an IRT model can bias the results in favor of IRT-based TEFs.

On the Results Observed

Our results generally agree with the findings from Sinharay and Holland (2010): regardless of IRT fit or data-generating process, most of the observed scenarios suggested that IRTOSE and IRTKE outperform KE with respect to bias and RMSE. Our preliminary tests have indicated, however, that much of the performance of KE seems to rely on how well the analyst can find a reasonable pre-smoothing model for the observed score distributions, particularly when the distribution contains scores with few observations. As a matter of fact, we have been able to substantially improve the performance of KE simply by choosing different pre-smoothing models, but not to the point where we could make that framework perform better or even as well as IRTOSE and IRTKE. Of particular interest is the occurrence of low-frequency scores, common on the extreme ends of the score range. The lack of data points on these intervals poses a problem for model fitting and is also a problem in large-scale assessments. The sensitivity of the equated scores to choices in model selection criteria and pre-smoothing model under the KE framework has also been observed by Wallin and Wiberg (2019). Finally, it should be noted that one must exercise care when figuring out how much "fine-tuning" should be done on a model, as too much of it could easily lead to a solution that overfits a model to a particular dataset, giving excessive and undue importance to characteristics that are just a product of chance.

The discussion on what the best model is should certainly involve more than the evaluation of a couple of statistical measures. For instance, KE can be quite useful when speed is a priority, since it does not require calculations at the item level and can offer results at a fraction of the time needed by IRT models. Moreover, even though IRT-based equating methods may not suffer from as high a degree of dependence on model fit as KE, they do require more assumptions to be implemented. As a matter of fact, it is pointless to try to fit a model if the data does not meet its basic assumptions in the first place. Thus, KE may offer a suitable alternative on instances where IRT should not be applied.

The continuity breaks observed on the bias and RMSE curves for the IRT data—see Figs. 3(b) and 4(b)—are caused by our decision to have all $\varphi(x)$ be integers for the IRT-generated data. These breaks occur at the points where the observed and the equated scores differ (by one unit, which explains the magnitude of the bias at those points), specifically at scores 11, 21, 36, 40, 59, 70. On Figure 2(b), these are also the scores where a break in the X-to-Y correspondence changes distance. For example, one can notice how score 11 on X—i.e., $x = 11$—is closer to score 12 on Y—$y = 12$—than to $y = 11$. This one-off difference continues until $x = 21$, where it is further expanded by one unit and then again at $x = 36$. From $x = 40$ onwards, the differences start unwinding until they reach 0. A third way to vizualize those differences is to notice how they tend to occur when the IRTOSE bias reaches 0 for scores between 0 and 40 or when it reaches -1 for scores larger than 40. Any method of continuization could smooth those breaks, but would have favored a particular framework. In any case, the presence of such breaks does not interfere with the readability of the figures, as separate parts of the plot can also be analyzed individually.

General Recommendations

It is important to keep in mind that the outcome of all the equating methods studied is a result of several choices of models and settings. Depending on the decisions taken at each step of the equating, we can observe variations in the output that could ultimately turn the decision in favor of a particular method and in detriment of another. The discussion about how to create the best environment possible to allow fair comparisons remains open, but we believe our contribution has helped shed some light into the debate.

Even if many consider score equating to be a "subjective art with theoretical foundations" (Cook & Paterson, 1987), we consider it worthy to pursue a truly fair comparison criterion that is as free from subjectivity as possible. This motivated our decision for a hands-free approach to pre-smoothing: instead of having a person manually checking the goodness-of-fit of countless models for over hundreds of samples, a stepwise procedure selected the model. The daily usage, however, often contains only one dataset and several methods to choose from. Under these conditions, we recommend attention to the framework assumptions, careful experimentation and observation of the sensitivity of the results to the different alternatives. This attention is especially important when dealing with high-stakes tests such as admissions tests, where the choice of a particular method can mean the difference between accepting an examinee into a university program or not.

Regarding the construction of a test booklet, we believe that focus should be put on the quality of the test items as well as on their quantity. Item quality is especially important if IRT models are expected to be used (due to their explicit reliance on item characteristics), although the efficacy of pre-smoothing methods for KE can also be harmed by the presence of items that are too difficult, easy, tricky or confusing.

Suggestions for Future Studies

Further contributions to this topic could focus not only on the application of the methods presented here to different data and other equating frameworks, but also their generalization to more than two test forms, internal anchor items and other equating frameworks such as those mentioned in *Introduction*. They could also analyze the sensitivity of the results to changes in the number of items, as well as explore alternative comparison methods, such as the one developed by Sinharay and Holland (2010) or other ranking-based methods.

It would also be interesting to see the development of other methods for working with real data; the method applied here was pointed out by Harris and Crouse (1993) as having some shortcomings such as the dependency on which form was taken as the base form—which we mitigated by randomly choosing the base form over hundreds of replications and averaging out the results—but they still consider it useful for checking the adequacy of an equating method or data collection design.

Some authors tend to use the term "circular equating" when equating a test form to itself, but we chose to avoid this nomenclature due to it not being universal. For example, Wang et al. (2000) found circular equating to be generally invalid for evaluating the adequacy of equating. However, they define circular equating as "equating a test form to itself through a chain of equating." A similar term, "equating in a circle," is used by Kolen and Brennan (2014, Section 8.4.2) to define a design where "Form X is equated to Form Y, Form Y is equated to Form Z, and Form Z is equated back to form X." The case studied in this paper differs from the former due to the absence of a chained structure, so their conclusions might not apply to the case under study. Still, they provoke further investigation into better ways to generate expected equated score distributions for empirical data, as well as studies comparing different equating transformations (e.g. linear equating and equipercentile equating) and covering test designs other than NEAT.

We chose well-known summary indices to evaluate the equating transformations, which leaves the creation of suitable equating-specific measures for future studies. For instance, the DTM can be easily applied to bias calculations, but more complex indices like the PRE might need adjustments before they can used in cross-framework comparisons. Future studies of these summary indices could also address other issues pointed out by Harris and Crouse (1993), such as the choice of the associated loss function.

A research paper addressing the comparison of equating transformations from different methods was published by Luecht and Ackerman (2018), discussing concepts like "truth" and "error" in IRT-based simulation studies. Their approach, however useful, is limited to the realm of IRT, and doesn't offer an ultimate solution to the arguably impossible task of defining a "truth" to be benchmarked against. Their insights could nonetheless provide a good starting point for an upgraded version of the methods proposed in the current manuscript.

Considering a different set of equating conditions might also offer extra insight on the results of this study. For example, one could control the differences in difficulty between the test forms, differences in examinee ability, correlation between anchor and main tests, and skewness.

Finally, we reinforce the importance of more research comparing different equating frameworks, even if it is unlikely that an unambiguous choice will surface from such studies (Kolen & Brennan, 2014, section 8.4.5) or even that a perfectly fair criterion can be found.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Fondazione Cassa di Risparmio di Padova e Rovigo, Swedish Wallenberg MMW (2019.0129), PRID, University of Udine (2017).

ORCID iDs

Waldir Leoncio https://orcid.org/0000-0002-6719-6162
Marie Wiberg https://orcid.org/0000-0001-5549-8262
Michela Battauz https://orcid.org/0000-0002-3098-689X

References

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

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, *82*(1), 48–66. https://doi.org/10.1007/s11336-016-9528-7

Battauz, M. (2015). EquateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*(7), 1–22. https://doi.org/10.18637/jss.v068.i07

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* (vol 1, pp. 9–49). Academic Press

Cook, L. L., & Paterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*(3), 225–244. https://doi.org/10.1177/014662168701100302

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT.* Technical Issues Related to the Introduction of the New SAT and PSAT/NMSQT (pp. 91–122).

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

González, J., Wiberg, M., & Von Davier, A. A. (2016). A note on the Poisson's binomial distribution in item response theory. *Applied Psychological Measurement*, *40*(4), 302–310. https://doi.org/10.1177/0146621616629380

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345–359. https://doi.org/10.1111/j.1745-3984.1990.tb00753.x

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*(3), 195–240. https://doi.org/10.1207/s15324818ame0603_3

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions. *ETS Research Report Series*, *1987*(2), i–40. https://doi.org/10.1002/j.2330-8516.1987.tb00235.x

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*(2), 133–183. https://doi.org/10.2307/1165330

Kim, K. Y., Lim, E., & Lee, W. C. (2019). A comparison of the relative performance of four irt models on equating passage-based tests. *International Journal of Testing*, *19*(3), 248–269. https://doi.org/10.1080/15305058.2018.1530239

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.) Springer

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional irt framework. *Mixed-format Tests: Psychometric Properties with a Primary Focus on equating*, *2*, 115–142

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. *Mixed-format Tests: Psychometric Properties with a Primary Focus on equating*, *2*, 13–44

Leôncio, W., & Wiberg, M. (2018). Evaluating equating transformations from different frameworks. In: *Springer proceedings in mathematics & statistics* (Vol 233). Springer

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*(3), 239–270. https://doi.org/10.1007/bf02289490

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*(2), 177–238. https://doi.org/10.1111/j.1745-3984.1977.tb00032.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ‘equatings. *Applied Psychological Measurement*, *8*(4), 453–461. https://doi.org/10.1177/014662168400800409

Luecht, R., & Ackerman, T. A. (2018). *A technical note on irt simulation studies: Dealing with truth, estimates, observed data, and residuals*. NCME

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing

Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the neat design. *Journal of Educational Measurement*, *47*(3), 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

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

Wallin, G., & Wiberg, M. (2019). Selecting a presmoothing model in kernel equating. In: *The annual meeting of the psychometric society* (pp. 111–120). Springer

Wang, T., & Brennan, R. L. (2009). A modified frequency estimation equating method for the common-item nonequivalent groups design. *Applied Psychological Measurement*, *33*(2), 118–132. https://doi.org/10.1177/0146621608314607

Wang, T., Hanson, B. A., & Harris, D. J. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement*, *24*(3), 195–210. https://doi.org/10.1177/01466210022031660

Wang, W., & Kolen, M. J. (2014). Comparison of the use of mc only and mixed-format common items in mixed-format test score equating. *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating*, *3*, 35

Wiberg, M., & González, 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