Smoothing of bivariate test score distributions – Model selection targeting test score equating

Abstract

Observed-score test equating is a vital part of every testing program, aiming to make test scores across test administrations comparable. Central to this process is the equating function, typically estimated by composing distribution functions of the scores to be equated. An integral part of this estimation is presmoothing, where statistical models are fit to observed score frequencies to mitigate sampling variability. This study evaluates the impact of commonly used model fit indices on bivariate presmoothing model-selection accuracy, in both item response theory (IRT) and non-IRT settings. It also introduces a new model-selection criterion that directly targets the equating function in contrast to existing methods. The study focuses on the framework of non-equivalent groups with anchor test design, estimating bivariate score distributions based on real and simulated data. Results show that the choice of presmoothing model and model fit index influence the equated scores. In non-IRT contexts, a combination of the proposed model-selection criterion and the Bayesian Information Criterion (BIC) exhibited superior performance, balancing bias and variance of the equated scores. For IRT models, high selection accuracy and minimal equating error were achieved across all scenarios.

Keywords: test score equating, smoothing, log-linear models, item response theory, model-selection.

1 Introduction

One of the key components for an educational testing program is to ensure fair assessments for all examinees. A routine task of most testing organisations therefore is test score equating, which refers to the procedure of putting scores from different test administrations on a common scale so that they can be compared and used interchangeably (González and Wiberg;

2017). The difficulty of a test may differ from time to time as many of the test items are typically not reused in order to ensure a high validity. Without any score adjustment, admission to for example university programs will therefore not be purely based on test-taker ability but also on the difficulty level of the test forms. Furthermore, in many realistic testing scenarios, the test groups are not randomized and therefore systematically differ from each other. In other words, the effect of test form difficulty is confounded by the latent ability, making it a highly non-trivial task to adjust for. The statistical task in test score equating therefore is to find a function that maps the scores of the new test form to the scale of the old, using samples from two non-randomized groups, thereby adjusting for the difference in difficulty while taking the difference in ability into account. Test score equating is thus a crucially important and fundamental component of fair assessments.

The starting point for any equating method is to define what notion of equivalence, in terms of the test scores and the latent ability, to use. The most common approach is to define two scores, x from test form X, and y from test form Y, as equivalent in terms of the latent ability measured by the test if they share the same relative position in their respective test score distribution. Following this definition, it is common practice to estimate the test score distributions by fitting parametric statistical models to the score data. Successfully applied, undesired irregularities of the observed score distributions are smoothed out, resulting in reduced sampling variability. This estimation is in the equating literature known as presmoothing. In this study, we view the test groups whose scores are to be equated as samples from two different ability populations. For this setting, most testing programs use a so called anchor test, meaning items which are common between test forms, to adjust for ability differences. For each respondent, we therefore have both a main test score and an anchor test score, and presmoothing of the observed test score distributions therefore involve bivariate data. This data collection design is commonly referred to as the nonequivalent groups with anchor test (NEAT) design. The equating function to be estimated is however defined only in terms of the main test score.

The most common presmoothing model is the log-linear model (Holland and Thayer;

1987; Kolen; 1991). There are several studies showing the positive effect of log-linear presmoothing on equating accuracy, see for example Hanson (1991), Livingston (1993), Moses and Holland (2007), and Moses and Liu (2011). Previous studies have also evaluated different model fit indices for log-linear models (Moses and Holland; 2010a) and their effect on certain traditional equating estimator (Moses and Holland; 2009; Liu and Kolen; 2020). There are furthermore other presmoothing options suggested in the literature, for example the beta-4 model (Kim et al.; 2005), the cubic B-spline and direct presmoothing (Cui and Kolen; 2009).

If the tests are calibrated with an item response theory (IRT) model, options for IRT equating are available as well. IRT equating methods have been implemented for both true-score and observed-score equating methods (Kolen and Brennan; 2014), thus considering both classical and modern test theory approaches. Recently, Andersson and Wiberg (2017) implemented IRT within kernel equating (von Davier et al.; 2004), an equating framework which includes both the traditional and modern equating methods.

This study is concerned with the sensitivity of presmoothing model-selection on the equated scores. We specifically evaluate the performance of the Akaike information criterion (AIC; Akaike 1974), the Bayesian information criterion (BIC; Schwarz 1978) and the Likelihood ratio chi-square statistic for selecting the parameterization of log-linear and IRT models, respectively, when the overall goal is test score equating. We furthermore propose a new model-selection method, which directly targets the equating function rather than the fit of the empirical score data. This study is different from Moses and Holland (2010a) since the selection of model parameterization is evaluated in terms of equated scores and not for the distributions being estimated. This study further differs from both Moses and Holland (2009) and Liu and Kolen (2020) who evaluate model fit indices for the equivalent groups (EG) design, as the focus here is on non-equivalent ability test groups. Moses and Holland (2010b) study model-selection for bivariate distributions for NEAT equating, but only considers the selection of cross-moments in the log-linear model and not the full model. They do moreover not consider IRT equating. In this study we furthermore considers a family

of equating functions which includes both traditional and modern equating methods, and non-IRT and IRT data. It makes this the most comprehensive study of model-selection for test score equating. The proposed model-selection criterion, which targets the asymptotic standard error of equating (ASEE; Holland and Thayer; 1989; von Davier et al.; 2004) and is based on the AIC, BIC and LRT, furthermore makes it possible to study the equating properties of these model-selection criteria. The results will be presented for both empirical data from a real admissions test and for a comprehensive simulation study.

The paper is organized as follows. We start by giving an introduction to test score equating and kernel equating, followed by a presentation of log-linear and IRT presmoothing. Next the empirical study is described, followed by the simulation study. The paper ends with a discussion and practical recommendations.

2 Test Score Equating

In this section, we will introduce all of the necessary definitions, notation and terminology. We begin with clearly defining the equating function, before presenting the kernel equating framework and the model-selection strategies that we consider in this study.

2.1 *Score Variables and Examinee Populations*

Let X and Y denote the test scores on test forms X and Y, respectively. The test scores X and Y are viewed as random variables with probability distributions as they are from randomly selected examinees from populations $\mathbf{P}$ and $\mathbf{Q}$, respectively. When the test groups are assumed to be randomly equivalent and administered test forms with different difficulty levels, equating is conducted using the EG design. This design only adjusts for differences in test form difficulty. If the samples however are assumed to be drawn from different populations, i.e., $\mathbf{P} \neq \mathbf{Q}$, the equating is preferably conducted using a NEAT design with a set of common items A as aid. To define equivalent scores, let $F_X(x) = \Pr(X \leq x|\mathbf{T})$ and $G_Y(y) = \Pr(Y \leq y|\mathbf{T})$, where $\mathbf{T}$ is the target population of the equating. Population $\mathbf{T}$ in

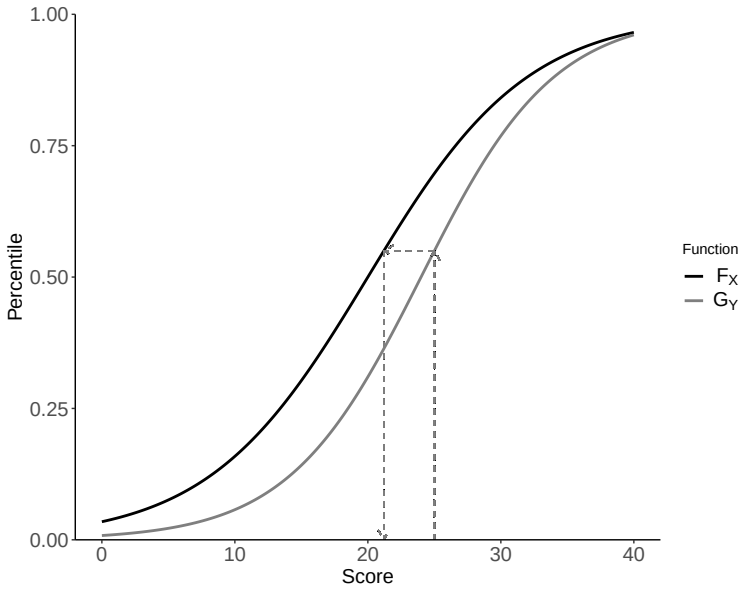


Figure 1: An illustration of the equipercentile transformation for two score distributions $F_X(x)$ and $G_Y(y)$. Two scores, x and y, are equivalent if $F_X(x) = p = G_Y(y)$, $p \in [0, 1]$.

the NEAT design could be thought of as a synthetic population and symbolically defined as $\mathbf{T} = w\mathbf{P} + (1 - w)\mathbf{Q}$, $w \in [0, 1]$. The term w determines how much weight is given to each population in the synthetic $\mathbf{T}$. As is custom, we let the relative sample sizes determine this when analysing simulated and real data. Assuming that X and Y are continuous, an equivalent score y on test form Y for a score x on test form X is obtained through the function $\varphi(x)$:

$$y = \varphi(x) = G_Y^{-1}\big(F_X(x)\big) \tag{1}$$

Equation 1 gives a general expression for the comparison of two continuous distributions (Wilk and Gnanadesikan; 1968) and is in the equating literature known as the equipercentile transformation (Kolen and Brennan; 2014; González and Wiberg; 2017). In Figure 1, the idea of the equipercentile transformation is illustrated. Note that the score variables need to be continuous for the equipercentile transformation to be properly defined. Since most testing programs utilize discrete test scores, certain continuous approximation is almost always required, which we present next.

2.2 *Test Score Equating Using Kernel Functions*

Test score equating can be described as comprising of four steps: 1) Estimation of the score probabilities, 2) Continuization, 3) Equating and 4) Evaluation of the equating function (von Davier et al.; 2004; González and Wiberg; 2017). Modern equating methods which employ a kernel function in the second step are commonly referred to as kernel equating methods (von Davier et al.; 2004). Since the first step is commonly combined with presmoothing of the score distributions, which we describe in the next section, we start with Step 2 here. To define the equating function within this framework, we begin by denoting the possible scores of X and Y by x_j, $j = 0, \ldots, J$, and y_k, $k = 0, \ldots, K$, respectively. Let $r_j = \Pr(X = x_j|\mathbf{T})$ and $s_k = \Pr(Y = y_k|\mathbf{T})$ denote the score probabilities on $\mathbf{T}$. For a population model, the continuization step in kernel equating approximates F_X with the cumulative distribution function (CDF) of the continuous random variable $\tilde{X} = a_X(X + h_X V) + (1 - a_X)\mu_X$, which equals

$$F_{\tilde{X}}(x) := P\big(\tilde{X} \leq x\big) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right), \tag{2}$$

where (μ_X, σ_X^2) are the mean and variance of X, respectively, $a_X = \sqrt{\sigma_X^2/\big(\sigma_X^2 + h_X\big)}$, $h_X > 0$ is a smoothing parameter, $V \sim \mathcal{N}(0, 1)$ and $\Phi(\cdot)$ is the CDF for V. The continuized score distribution $G_{\tilde{Y}}$ of the score variable $\tilde{Y}$ is obtained in a similar way.

The bandwidth h_X, which determines the smoothness level of the continuous approximation to F_X, can be selected in several ways but the equated scores have been shown to be robust to different choices (Wallin et al.; 2021). In this paper, we adopt the common approach of selecting the bandwidth which minimizes

$$Q(h_X) = \sum_j \big(\hat{r}_j - \frac{d}{dx}\hat{F}_{\tilde{X}}(x)\big)^2. \tag{3}$$

Two popular methods for equating in the NEAT design are Chained Equating (CE) and

Post-Stratification Equating (PSE). The CE function is given by

$$\varphi_{Y(CE)}(x) = G^{-1}_{Q\tilde{Y}}(H_{Q\tilde{A}}(H^{-1}_{P\tilde{A}}(F_{P\tilde{X}}(x)))) \tag{4}$$

where $F_{P\tilde{X}}$, $H_{P\tilde{A}}$, $H_{Q\tilde{A}}$ and $G_{Q\tilde{Y}}$ are continuous approximations of the type in (2) to the CDFs of X and A on population $\mathbf{P}$ and Y and A on population $\mathbf{Q}$, respectively. Under a population invariance assumption of the two links in (4), it can be shown that the CE estimator aligns with the equipercentile transformation in (1).

The PSE function is given by

$$\varphi_{Y(PSE)}(x) = G^{-1}_{T\tilde{Y}}\big(F_{T\tilde{X}}(x)\big), \tag{5}$$

where $F_{T\tilde{X}}$ and $G_{T\tilde{Y}}$ are continuous approximations of F_X and G_Y on the target population $\mathbf{T}$.

We conclude this section by pointing out that, with the bandwidth as only exception, the distributions $F_{\tilde{X}}$, $G_{\tilde{Y}}$, and $H_{\tilde{A}}$ are entirely dependent on their means and standard deviations, all of which are functions of the population test score probabilities. This paper is consequently entirely focused on the estimation of these probabilities.

3 Presmoothing of Test Score Distributions

In this section, we present two ways of estimating the population test score probabilities, depending on whether an IRT or non-IRT approach is used. We present the two most common model choices and begin with the non-IRT approach, for which log-linear model are the most common presmoothing model.

3.1 *Log-Linear Models*

Let the possible score values of the anchor score A be denoted by a_l, $l = 1, \ldots, L$, where L is the number of binary scored anchor items. Let n_{jl} denote the number of examinees scoring $X = x_j$ and $A = a_l$, with $\sum_{j,k} n_{jl} = N$, and m_{kl} denote the number of examinees

scoring $Y = y_k$ and $A = a_l$, with $\sum_{k,l} m_{kl} = M$. Further let $\mathbf{p}$ and $\mathbf{q}$ denote the respective probability vectors for counts $n_{11}, \ldots, n_{JL}$ and $m_{11}, \ldots, m_{KL}$. Assume that

$$\mathbf{n} = (n_{11}, \ldots, n_{JL})^\top \sim \text{Multinomial}(N, \mathbf{p}),$$

$$\mathbf{m} = (m_{11}, \ldots, m_{KL})^\top \sim \text{Multinomial}(M, \mathbf{q}),$$

and that $\mathbf{n}$ and $\mathbf{m}$ are independent from each other. From here on, only formulas for the (X, A) scores will be presented, and the analogous expressions for (Y, A) will be suppressed.

In the NEAT design, the bivariate distributions of (X, A) and (Y, A), respectively, are estimated. Following from the assumptions stated above, the log-linear model for the bivariate distribution of (X, A) equals

$$\log(p_{jl}) = \beta_0 + \sum_{i=1}^{I} \beta_{x,i} x_j^i + \sum_{b=1}^{B} \beta_{a,b} a_l^b + \sum_{o=1}^{O} \sum_{e=1}^{E} \beta_{xa,de} x_j^d a_l^e, \tag{6}$$

where $p_{jl} = \Pr(X = x_j, A = a_l)$, β_0 is a normalizing constant, the β:s are parameters that need to be estimated, and x_j^i and a_l^h are functions of the score variables.

As for the univariate case, the bivariate log-linear model possesses a moment-matching property when the parameters are estimated using maximum likelihood (Moses and Holland; 2010a). From (6) it follows that I and B sample moments in the marginal distributions of X and A, respectively, are preserved, and O and E determine the number of observed bivariate moments that are preserved. We illustrate this property in Figure 2, where we have generated univariate data for which we fit a sequence of log-linear models. For each window in the figure, we include one additional polynomial term, thus preserving one additional moment in the test score distribution.

Once we have fitted a log-linear model to (X, A) and (Y, A), we need functions that turn the estimated joint probabilities $\hat{p}_{jl} = \widehat{\Pr}(X = x_j, A = a_l)$ and $\hat{q}_{kl} = \widehat{\Pr}(Y = y_k, A = a_l)$ into the marginal probabilities required by the CE and PSE methods. For CE, we need the marginal probabilities of X, Y and A in populations $\mathbf{P}$ and $\mathbf{Q}$ to plug into their continuized

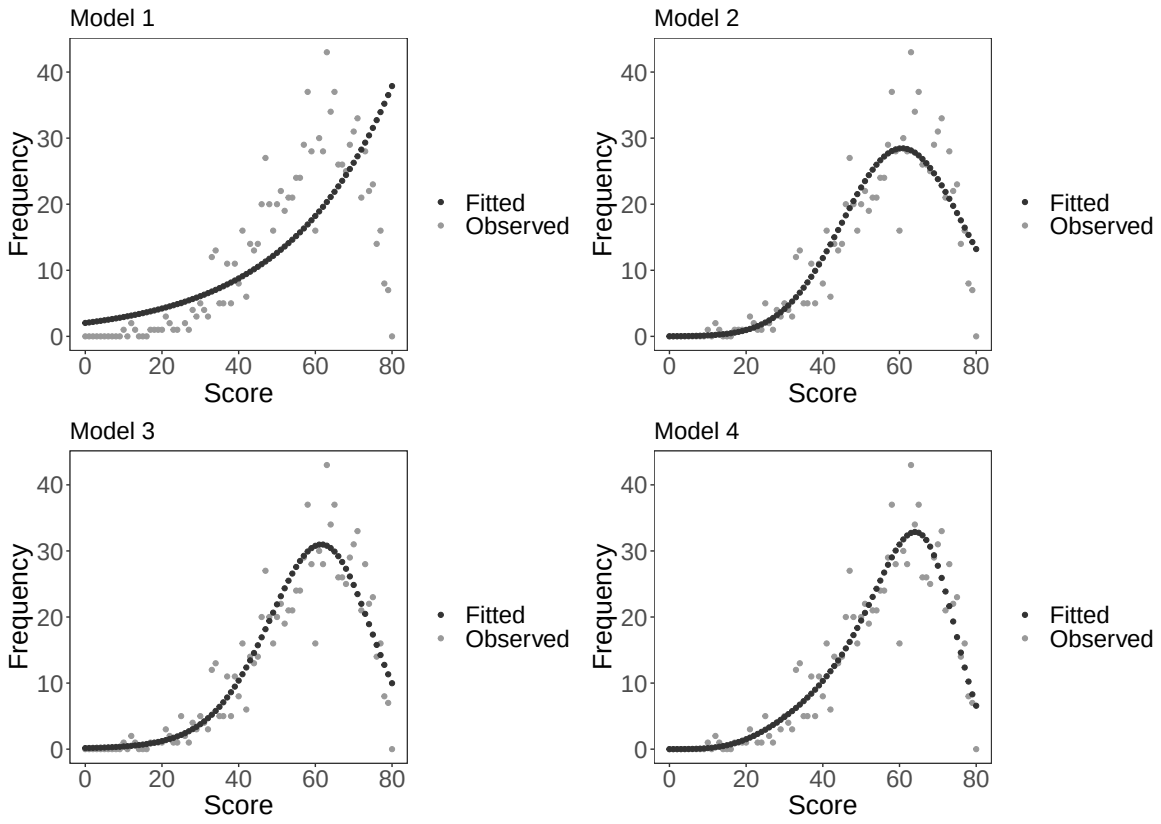


Figure 2: Log-linear models with $I = 1$ to $I = 4$.

CDFs. We get these by

$$\hat{r}_{Pj} = \widehat{\Pr}(X = x_j|\mathbf{P}) = \sum_l \hat{p}_{jl},$$

$$\hat{t}_{Pl} = \widehat{\Pr}(A = a_l|\mathbf{P}) = \sum_j \hat{p}_{jl},$$

$$\hat{t}_{Qj} = \widehat{\Pr}(A = a_l|\mathbf{Q}) = \sum_k \hat{q}_{kl},$$

$$\hat{s}_{Qj} = \widehat{\Pr}(Y = y_k|\mathbf{Q}) = \sum_l \hat{q}_{kl}$$

For PSE, we transform $\hat{\mathbf{p}} = (\hat{p}_{11}, ..., \hat{p}_{JL})^\top$ and $\hat{\mathbf{q}} = (\hat{q}_{11}, ..., \hat{q}_{KL})$ into $\hat{\mathbf{r}} = \widehat{\Pr}(X = x_j|\mathbf{T})$ and $\hat{\mathbf{s}} = \widehat{\Pr}(Y = y_k|\mathbf{T})$ by

$$\hat{\mathbf{r}} = \sum_l \left[w + \frac{(1-w)\hat{t}_{Ql}}{\hat{t}_{Pl}} \right] \hat{\mathbf{p}}_l$$

and

$$\hat{\mathbf{s}} = \sum_l \left[(1-w) + \frac{w\hat{t}_{Pl}}{\hat{t}_{Ql}} \right] \hat{\mathbf{q}}_l,$$

where $\hat{\mathbf{p}}_l = \begin{bmatrix} \hat{p}_{1l} & \cdots & \hat{p}_{Jl} \end{bmatrix}^\top$ and $\hat{\mathbf{q}}_l = \begin{bmatrix} \hat{q}_{1l} & \cdots & \hat{q}_{Kl} \end{bmatrix}^\top$.

3.2 *IRT Models*

Another approach for estimating the score probabilities within the kernel equating framework is by using an IRT model (Andersson and Wiberg; 2017). For this class of models, the underlying assumption is that the probability of answering an item correctly is a function of the latent ability and of item parameters that determines the shape of that function. Let $X_{ij^*} = 1$ if examinee i from population $\mathbf{P}$ answers item $j^* \in \{1, 2, ..., x_J\}$ correctly, and $X_{ij^*} = 0$ otherwise. Let p_{ij^*} denote the probability that $X_{ij^*} = 1$. Let $\theta_i \in \{-\infty, \infty\}$ denote the latent ability of test-taker i, $\alpha_{j^*} \in [0, \infty\}$ the discrimination of item j^*, $b_{j^*} \in \{-\infty, \infty\}$ the difficulty of item j^*, and $c_{j^*} \in [0, 1]$ the lower asymptote, or guessing parameter. If the three-parameter logistic (3-PL) model is used to calibrate the item pool, the probability p_{ij^*} equals

$$p_{ij^*} = \Pr(X_{ij^*} = 1 | \alpha_{j^*}, b_{j^*}, c_{j^*}; \theta_i) = c_{j^*} + \frac{1 - c_{j^*}}{1 + \exp\big(-\alpha_{j^*}(\theta_i - b_{j^*}) \big)} \tag{7}$$

When $c_{ij^*} = 0$, the model in Equation 7 reduces to the two-parameter logistic (2-PL) model and when, additionally, $\alpha_{j^*} = 1$, the one-parameter (1-PL), or Rasch, model is obtained (Hambleton and Swaminathan; 1985).

As the probability specified in Equation 7 is on item-level and not for the sum-score, the estimated probabilities need to be cumulated. Therefore, let

$$X_i = \sum_{j^*=1}^{x_J} X_{ij^*}$$

denote the sum-score of individual i. The probability distribution of X_i for a given ability

θ_i is given by the compound binomial distribution (Birnbaum; 1968),

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

As the compound binomial distribution is computationally intensive (González et al.; 2016) an iterative process such as the Lord and Wingersky (1984) algorithm can be used to calculate the probabilities. See Andersson and Wiberg (2017) for explicit approximating formulas to generate the score probabilities for CE and PSE for IRT equating.

4 Goodness-of-fit indices

Model fit indices are often classified as belonging to either a significance testing strategy or a parsimony strategy (Moses and Holland; 2010a; Liu and Kolen; 2020). For the former, there have been several statistics suggested which are asymptotically chi-square distributed. One option is the the likelihood ratio test (LRT; Haberman; 1974a,b), which is based on the following test statistic:

$$2 \sum_j n_{jl} \log \left(\frac{n_{jl}}{N \hat{p}_{jl}} \right),$$

where n_{jl} and $N\hat{p}_{jl}$ are the observed and estimated frequencies, respectively, of $\{X = x_j, A = a_l\}$. Other common choices are the Cressie-Read chi-square (Read and Cressie; 1988), the Pearson chi-square and the Freeman-Tukey chi-square (Holland and Thayer; 2000). For nested models such as log-linear models and the 1-PL, 2-PL and 3-PL IRT models, it is possible to compare the fit of a complex model relative to that of a simpler model by calculating the probability of such difference as measured by their respective chi-square statistics. This discrepancy measure gets its p-value from the chi-square distribution with degrees of freedom equal to the difference in the number of parameters of the models.

The indices belonging to the parsimony strategy try to balance model fit with the parameterization of the model. The model selection is made by comparing a number of competing models with different parameterizations with respect to some suitable statistic. Two common

choices for this statistic are the AIC and the BIC, which are defined as

$$\text{AIC} = -2\log(L_d) + 2d,$$

$$\text{BIC} = -2\log(L_d) + \log(N)d,$$

where $\log(L_d)$ denotes the log-likelihood function for $\mathbf{p}$, and d equals the number of parameters being estimated. The term $-2\log(L_d)$ equals the likelihood ratio chi-square statistic, and as such, the functional form differs between the considered models. For example, for a bivariate log-linear model of (X, A),

$$\log(L_d) = \sum_{jl} n_{jl} \log(\hat{p}_{jl}),$$

and for the 2-PL model,

$$\log(L_d) = \frac{\exp\left\{\theta_i \sum_{j^*} x_{ij^*} a_{j^*} - \sum_{j^*} x_{ij^*} a_{j^*} \beta_{j^*}\right\}}{\prod_{j^*}\left[1 + \exp\left\{a_{j^*}(\theta_i + \beta_{j^*})\right\}\right]}.$$

In this study, the same approach as in Liu and Kolen (2020) is taken, where only the likelihood ratio chi-square, AIC and BIC indices are evaluated. As pointed out in Liu and Kolen (2020), previous studies such as Moses and Holland (2009, 2010a) have shown that these indices performs as well or better than the other exisisting indices. For the significance testing strategy, a significance level of $1-(1-\alpha)^{1/(\#Models-1)}$ is used for the individual tests, where $\#Models$ denotes the total number of models that are tested.

Recently, Brown et al. (2015) pointed out that although the 1-PL, 2-PL and 3-PL models are nested, the likelihood ratio test is not appropriate when selecting between the 2-PL and 3-PL model. This is due to the fact that the guessing parameter c_{j^*} for such test is set to its boundary value ($c_{j^*} = 0$) in the null hypothesis. This violates one of the assumptions of the likelihood ratio test, making the chi-square distribution as a reference distribution invalid. In the literature of mixed-effects modeling however, the problem of testing a null hypothesis that is on the boundary of the parameter space is well-known. Brown et al. (2015) suggest to

use the p-values from a simulated null distribution, as given by the **ltm** package (Rizopoulos; 2006) of the statistical software R (R Core Team; 2022), when using the likelihood ratio test to determine whether the guessing parameter is significantly different from zero. This is therefore the approach taken in this study.

4.1 Our Implementation

In our implementation, we set up candidate models up to the user-specified polynomial degree for the univariate and cross moments, i.e. the I, B, O, and E terms in (6). Since this creates a very large set of possible models, we adopt a two-step approach. In the first step, we let each criterion select the best-fitting model among the set of models with only univariate moments, not including any cross-moments. In the second step, we take the best-fitting models from the first step according to each respective criterion and sequentially add cross-moments. The final model for each criterion is selected from this set, and still allows for the model without any cross-moments, i.e. the best-fitting model from the first step, to be the final model chosen. For each pair of selected models, the test forms are equated, resulting in one estimated equating function per model-selection criterion.

4.2 *A Selection Algorithm Targeting the Equating Function*

We propose a new model selection algorithm for log-linear presmoothing that directly targets the equating function. The idea is a two-step approach: First, we search for the best-fitting models to the bivariate data (X, A) and (Y, A), according the AIC, BIC and LRT, respectively. This will give us at most six unique models, one (X, A)-model and one (Y, A)-model for each criterion. In the second step, we search through all combinations of model pairs, and select the pair which minimizes certain loss function which is related to the equating function. For this function, we propose the ASEE. Based on the asymptotic normality of the maximum likelihood estimator of $\begin{bmatrix}\hat{\mathbf{r}} & \hat{\mathbf{s}}\end{bmatrix}^\top$, the asymptotic distribution of the equating

function is given by

$$\varphi(x;\hat{\mathbf{r}},\hat{\mathbf{s}}) \xrightarrow{\mathrm{d}} \mathcal{N}\big(\varphi(x;\mathbf{r},\mathbf{s}), \mathbf{J}_{\varphi_Y}\mathbf{J}_{\mathrm{DF}}\mathbf{C}\mathbf{C}'\mathbf{J}'_{\mathrm{DF}}\mathbf{J}'_{\varphi_Y}\big)$$

for all x (von Davier et al.; 2004; Holland and Thayer; 1989). Here, $\mathbf{J}_{\varphi_Y}$ is the Jacobian of the equating function, $\mathbf{J}_{\mathrm{DF}}$ represents the Jacobian of the design function, and $\mathbf{C}$ is connected to the score distributions' covariance matrix. For a detailed derivation of the ASEE, refer to von Davier et al. (2004). The ASEE is consequently given by

$$\mathrm{ASEE}(x) = ||\mathbf{J}_{\varphi_Y}\mathbf{J}_{\mathrm{DF}}\mathbf{C}||, \tag{8}$$

where $||\cdot||$ denotes the Euclidean norm. We summarize the procedure in Algorithm 1.

Algorithm 1 Log-linear model selection targeting test score equating

Input: Observed bivariate data (X_i, A_i), $i = 1, ..., N_P$ and (Y_i, A_i), $i = 1, ..., N_Q$, where N_P and N_Q are the respective sample sizes of the test groups.

1. Fit log-linear models to the bivariate data according to (6), for an increasing number of univariate and cross-moments. Select the best (X, A) and (Y, A) models according to the AIC, BIC and LRT, respectively.
2. Search among the selected models in Step 1 to find the pair that minimises

$$\frac{1}{J+1}\sum_{j=0}^{J}\mathrm{ASEE}(x_j) \tag{9}$$

where J equals the test length, and $\mathrm{ASEE}(x_j)$ denotes the ASEE for the x_jth test score.

Output: One fitted log-linear model for (X, A) and one for (Y, A).

It is possible to give different importance to score points. In that case, $1/(J+1)$ in (9) can be replaced with other weights δ_j such that $\sum_{j=0}^{J}\delta_j = 1$, so that each δ_j conveys importance to score x_j. The criterion then becomes $\sum_{j=0}^{J}\delta_j\mathrm{ASEE}(x_j)$.

Note that Algorithm 1 can be seen as both a way to select the log-linear presmoothing models but also as a way to evaluate the AIC, BIC and LRT criteria when the goal is to equate test forms. By simulating test data and selecting the presmoothing model according to the AIC, BIC and LRT, respectively, we can count how many times Algorithm 1 selects each criterion. In that way we can use Algorithm 1 to evaluate which of the AIC, BIC and

LRT most often select models that minimize the ASEE. Since the output of Algorithm 1 is a pair of presmoothing models, we are in turn able to evaluate Algorithm 1 as a model-selection tool. Note also that this algorithm does not make sense to use in an IRT setting since we do not wish to end up with for example a Rasch model for the (X, A) data and a 3-PL model for the (Y, A) data. From this point forward, we will refer to this model-selection approach as the ASEEmin CE or ASEEmin PSE depending on whether CE or PSE is used. We lastly point out that the term δ_j is used as a weight to give, possibly, different importance to scores. In certain testing situations, score points near certain cut-scores might convey more importance and it might therefore be better if ASEEs are smaller around cut-scores than they are for other score-points[1]. In this paper we only consider $\delta_j = 0, j, ..., J$.

5 Empirical Study

In this section, we explore the influence of presmoothing techniques and model selection on the equated scores, using real data from two forms of the Swedish Scholastic Aptitude Test (SweSAT). The test comprises both a verbal and a quantitative section, each of them equated separately. Each section consists of 80 binary-scored multiple-choice items. In addition, a 40-item anchor test is administered. For this study, we utilized a sample of 7,322 examinees from each of two groups from the quantitative section for a test administration within the last ten years. The new test form is denoted as X and the previous form as Y. Detailed information about this quantitative test form and its associated anchor test can be found in Table 1.

For instance, a clear difference between the mean scores of the new and old test forms and their respective skewness can be clearly observed. The correlation between the new test form (X) and the anchor test scores stands at 0.808, while that between the old test form (Y) and the anchor test scores is slightly lower at 0.806.

Figure 3 displays the score distributions for both X and Y, in addition to the anchor test scores, denoted as A_X and A_Y. A noticeable difference in the distributions is evident,

[1]We thank one of the anonymous reviewers for this idea and motivation.

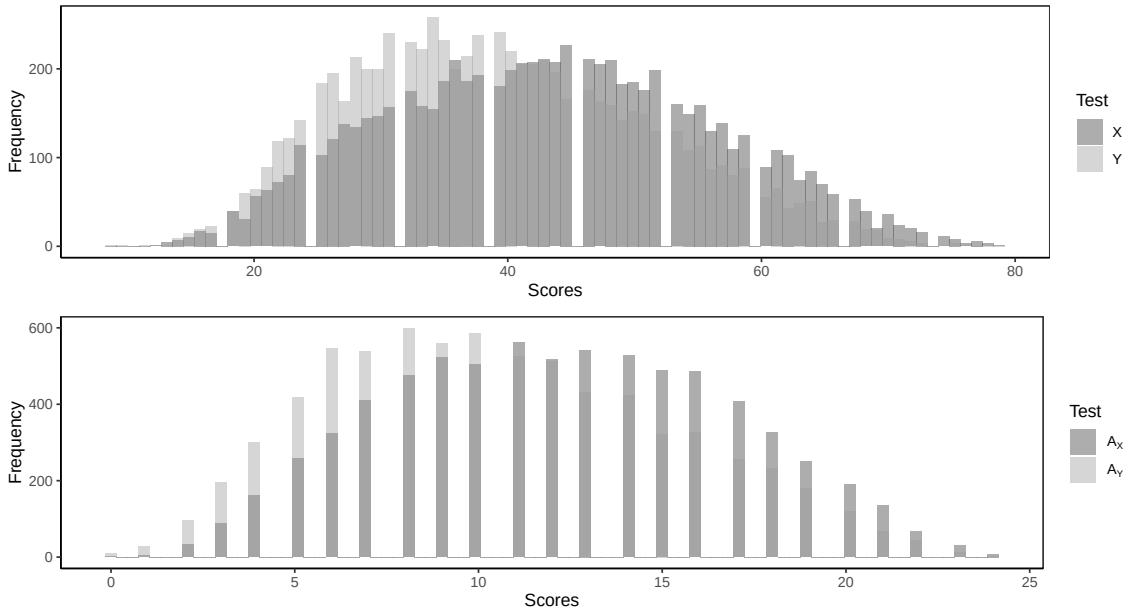


Figure 3: The score distributions of the X and Y scores (upper panel) and the anchor scores (lower panel), where X represents the new test form and Y the old test form.

indicating either a higher average ability level among respondents taking the new test form X, an easier test form, or possibly both. The distributions of the anchor test scores suggest that a difference in ability between the groups is present.

At the time of these test administrations, a SweSAT result retained its validity for five years. Examinees were permitted unlimited attempts, with only the highest test result being considered when applying to a university program. In practice, the SweSAT is equated using non-IRT methods under the NEAT design (Lyrén and Hambleton; 2011). The available data consists solely of the total score, as opposed to individual item scores. Therefore, we will employ log-linear presmoothing and equate the test forms using the smoothed score distributions for both CE and PSE.

Table 1: Mean, standard deviation (SD), skewness and number of examinees (N) in the empirical study and for the created test forms used in the simulation study.

Test form	Mean	SD	Skewness
X	43.315	12.654	0.098
Y	39.341	11.802	0.336
A_X	12.171	4.593	0.063
A_Y	10.555	4.643	0.303

Table 2: The selected log-linear models for each model fit index.

Model	AIC	BIC	LRT	ASEEmin CE	ASEEmin PSE
(X, A)	X^5, A^4, $X^3 : A^3$	X^5, A^4, $X^3 : A$	X^5, A^5, $X^3 : A^3$	X^5, A^4, $X^3 : A$	X^5, A^4, $X^3 : A$
(Y, A)	Y^5, A^4, $Y^3 : A^3$	Y^5, A^4, $Y : A$	Y^5, A^4, $X^3 : A^3$	Y^4, A^4, $Y : A$	Y^4, A^4, $Y : A$

5.1 *Results of Empirical Study*

Table 2 shows the presmoothing models chosen by each model-selection criterion. Only the highest power moment is displayed, meaning for example that X^2 implies that both X and X^2 are included in the model, and $X^3 : A^3$ means that the interactions $X : A$, $X : A^2$, $X : A^3$, $X^2 : A$, $X^2 : A^2$, $X^2 : A^3$, $X^3 : A$, $X^3 : A^2$ and $X^3 : A^3$ are included.

We see that the BIC, ASEEmin CE and ASEEmin PSE select the same (X, A)-model, i.e., the BIC selects the model that minimizes the ASEE when either CE or PSE is used to equate the test forms. For the (Y, A)-model, the BIC includes the fifth univariate moment for Y, which the ASEEmin criteria do not do. The AIC and LRT select nearly identical models, which include more cross-moments than the BIC and ASEEmin criteria does. Noticeably, all of the traditional criteria select slightly different models. The next question is to determine to what extent this affects the equated scores and the corresponding ASEE values.

Figure 4 demonstrates the estimated equated scores (with the raw score deducted) and the corresponsing ASEE values by the CE (left column) and PSE (right column) estimator. Due to the selection of identical pairs of presmoothing models, the BIC and ASEEmin criteria produce identical results. Since the AIC and LRT criteria selected very similar models, their results are also close to identical. In Figure 4 we therefore see two unique curve patterns in each panel: the solid line, derived from models chosen by the AIC, and the dot-dashed line, which is the result of presmoothing models chosen via the BIC and ASEEmin.

These curves reveal that the equated scores and ASEE values share a high degree of similarity across a substantial portion of the score scale. At the extremes of the score scale, however, differences emerge. These differences at the tails exceed the so-called 'difference that matters' (DTM; Dorans and Feigenbaum; 1994), defined as a difference of half a score unit. In practical terms this indicates that these methods will produce different equated

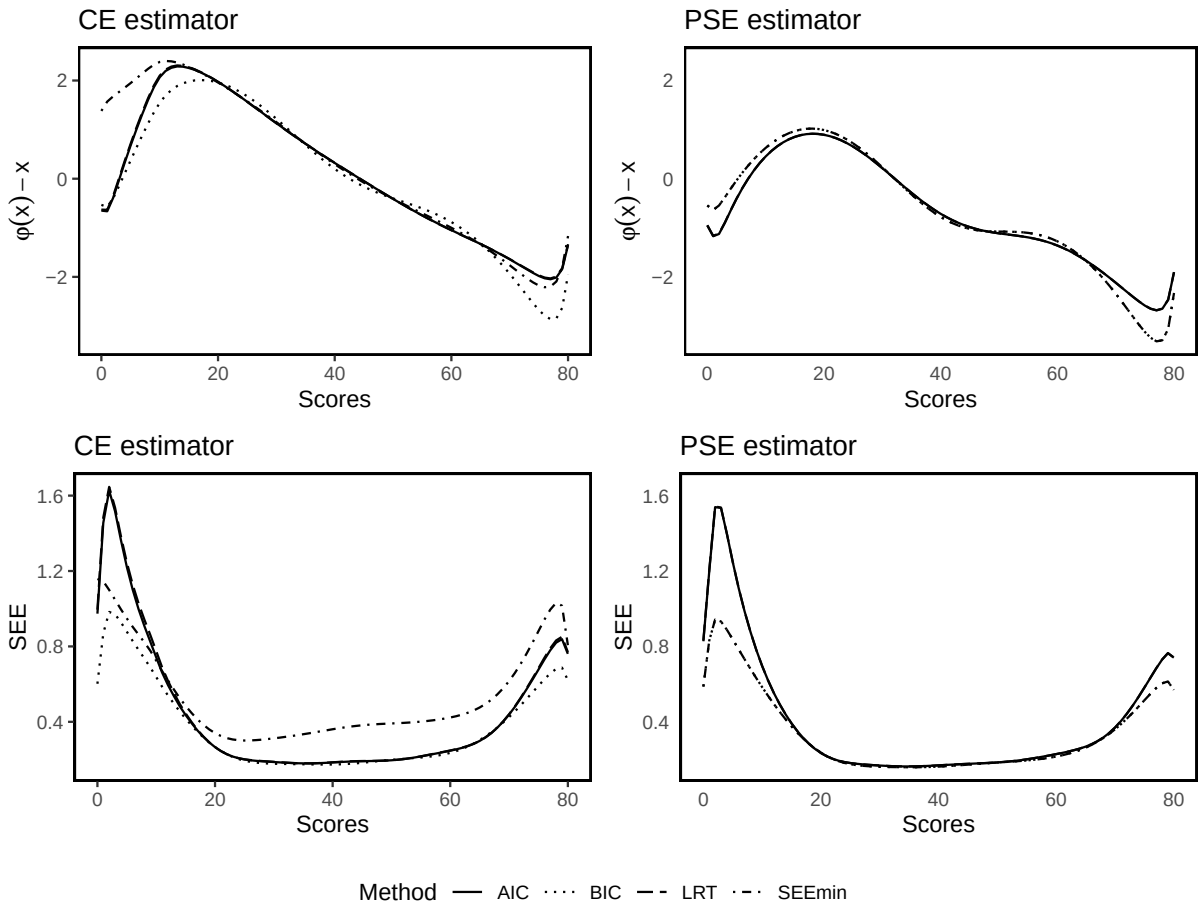


Figure 4: The difference between the equated scores and the raw scores (first row) and the ASEEs (second row), for each respective equating estimator using the CE and PSE estimator, respectively.

scores at the score range where many critical decisions are being made. Therefore, the equated scores depend significantly on the presmoothing model-selection criterion. Lastly, we note that the BIC and ASEEmin criteria produce substantially smaller ASEE values for the lowest and highest score values.

From the empirical analysis, we conclude that the presmoothing model seletion have an impact on the equated scores, and the magnitude of this influence is dependent on the choice of equating estimator. We study this phenomenon in greater depth in the following section, which includes a comprehensive simulation study.

6 Simulation Study

We evaluate the performance of all of the considered model-selection criteria and their influence on the estimation of the equating function through an extensive simulation study. We consider two setups that will be treated in separate subsections. The first setup generates data from an underlying log-linear model and the second setup from an underlying IRT model. In this way, we can evaluate the performance of the model selection criteria both when the score distributions are smoothed using log-linear models, and when the test is calibrated with an IRT model. For both setups, we vary the underlying, true data-generating process, together with other key quantities such as the sample size and the number of items. The results are based on 100 iterations using sample sizes of 3,000 and 6,000, and with test lengths of 40 and 80. When the test length is 40, we set the anchor length to 13, and when the test length is 80, we set it to 25. Since the results for a sample size of 3,000 and of 6,000 were very similar, we have placed the results based on the larger sample size in the Appendix. For the same reason, we did the same for some of the results based on $J = 80$. Most of the computations are made using the R package **kequate** (Andersson et al.; 2013), which includes functions for conducting test score equating using kernel functions. We also use the R package **copula** (Yan; 2007) to generate bivariate data. Our own functions, including the proposed algorithm for model-selection, have been included with the submission of this manuscript and will be made publicly available as an R package upon acceptance of the paper.

6.1 *Setup A - Data Generated by Log-Linear Models*

In the first setup, we follow the design of Wallin et al. (2021). We therefore start by generating the true score probabilities $p_{jl} = \Pr(X = x_j, A = a_l)$, $j = 0, \ldots, J, l = 0, \ldots, L$ and $q_{kl} = \Pr(Y = y_k, A = a_l)$, $k = 0, \ldots, K, l = 0, \ldots, L$. In the following, we describe only the procedure for the (X, A) probabilities since the (Y, A) probabilities are generated in the same way. To generate p_{jl}, we begin by generating auxiliary score variables (U_i, V_i), $i = 1, ..., N$,

using a normal copula bivariate distribution, which is a joint distribution for (U, V) and where we set the marginal distributions to follow Beta(α, β) distributions. The shape parameters α and β are set according to three distributional settings, as soon explained. The correlation between U and V is set to 0.82 to mimic the strong correlation often seen in real data, such as in the Empirical Illustration. After the auxiliary scores have been generated, score variables X and Y are generated as $X_i = \lfloor (J-1)U_i \rceil$ and $A_i = \lfloor (L-1)U_i \rceil$, meaning that the auxiliary variables are multiplied by the test length and thereafter rounded to the nearest integer so that they are discrete random variables with support in the range $(0, J)$ and $(0, L)$, respectively. In this simulation study, we consider $J = \{40, 80\}$ and $L = \{13, 25\}$. Next, a log-linear model is fit to the (X, A) data. The model fits the fourth power of X and the third power for A, which is based on visual inspection. The estimated probabilities from the model fitted to the population-level data are now treated as the true score probabilities $p_{jl} = \Pr(X = x_j, A = a_l)$.

In the next step, we sample test score frequencies $\boldsymbol{n}_{JL} = \{n_{jl} = \sum_{i=1}^{N} I(X_i = x_j, A_i = a_l)\}$, $j = 1, \ldots, J, l = 1, \ldots, L$, as $\boldsymbol{n}_{JL} \sim$ Multinomial$(N, (p_{11}, \ldots, p_{JL}))$. The model selection criteria are thereafter employed to select the most fitting log-linear models for the (X, A) and (Y, A) combinations, considering models up to the sixth polynomial degree for the univariate moments, and up to the third cross-moment, i.e. $I = B = \{1, 2, 3, 4, 5, 6\}$, and $O = E = \{1, 2, 3\}$ in (6). Since this creates a very large set of possible models, we adopted a two-step approach. In the first step, we let each criterion select the best-fitting model among the set of models with only univariate moments, not including any cross-moments. In the second step, we take the best-fitting models from the first step according to each respective criterion and sequentially add cross-moments. The final model for each criterion is selected from this set, and still allows for the model without any cross-moments, i.e. the best-fitting model from the first step, to be the final model chosen. For each pair of selected models, the test forms are equated, resulting in one estimated equating function per model-selection criterion.

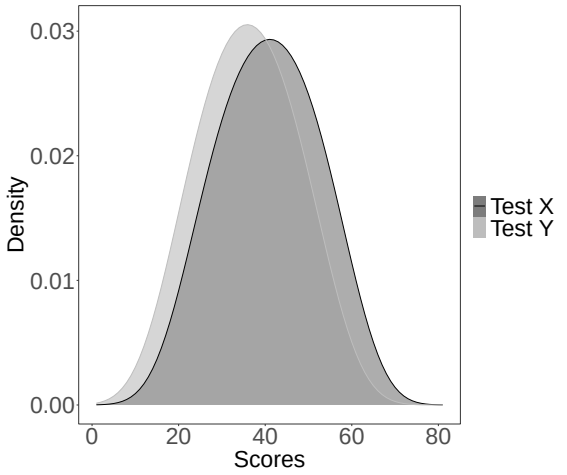
In Figure 5 we illustrate the true score distributions $r_j = \Pr(X = x_j|\mathbf{T})$, $j = 0, ..., J$, and $s_k = \Pr(Y = y_k|\mathbf{T})$, $k = 0, ..., K$ for each scenario, considering the shape parameters

for the beta distribution to be $(\alpha = 5, \beta = 5)$ for a symmetric setting, $(\alpha = 5, \beta = 2)$ for a skewed setting and to produce a bimodal setting, a mixture of Beta distributions with $(\alpha = 25, \beta = 15)$ and $(\alpha = 15, \beta = 25)$ is used. Note that for the Y data, we shift the data by 2 units along the score axis, making it represent a more difficult test. Note that with the true r_j and s_k generated, we can define the population-level score CDFs in Equation (2), and thus, the true equating transformations in Equations (4) and (5), respectively. When we are generating our samples, we are therefore sampling test scores from the true test score probability distributions illustrated in Figure 5. The anchor score distributions, illustrating the ability differences, are shown in Figure 6.

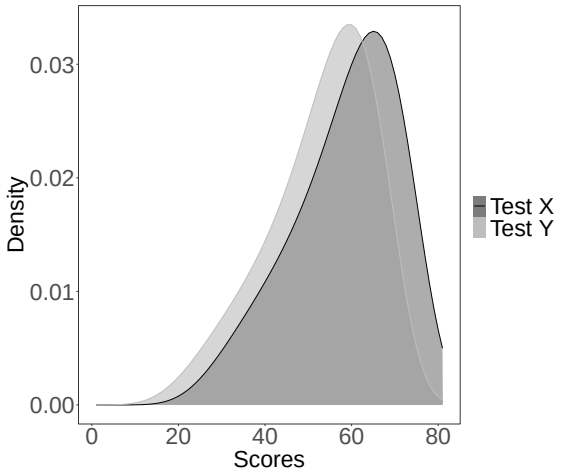
6.2 *Set-Up B – Data Generated by IRT Models*

In Set-Up B, the data is generated by an IRT model, where we let the true model equal the Rasch, 2-PL and 3-PL model, respectively. The data generation and equating estimation procedure follows the kernel equating IRT method proposed in Andersson and Wiberg (2017). This equating method generates test score probabilities from IRT models using the Lord-Wingersky algorithm (Lord and Wingersky; 1984) and use these probabilities in the continuization step of the kernel equating framework. First, we generate item parameters for the test forms. These parameters include difficulty parameters, discrimination parameters and guessing parameters, depending on which model defines the true model. The difficulty parameters are generated from a $\mathcal{N}(0, 1)$ distribution, the discrimination parameters are drawn from a Uniform(0.52) distribution and the guessing parameters are drawn from a $\mathcal{N}(0.25, 0.5)$ distribution.

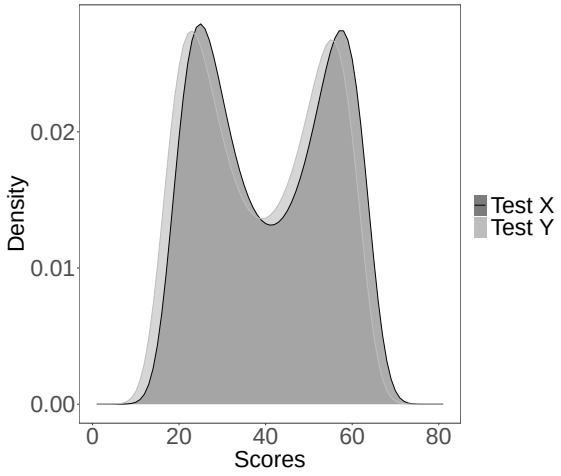
For each individual in the $\mathbf{P}$ test group, an ability is drawn from a standard normal distribution and for each individual in the $\mathbf{Q}$ group, an ability is drawn from a $\mathcal{N}(0.5, 1.2)$ distribution. Based on this ability and the previously defined item parameters, the individual's responses to the items on the test forms and anchor test are generated through the item response function in (7). The final dataset is obtained by randomly selecting a sample from the population for each form. Thereafter, a model selection process begin where the choice



(a) True test score probabilities generated from symmetric distributions.



(b) True test score probabilities generated from skewed distributions.

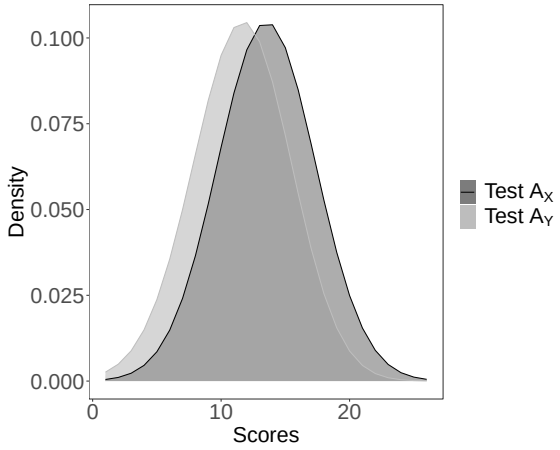


(c) True test score probabilities generated from bimodal distributions.

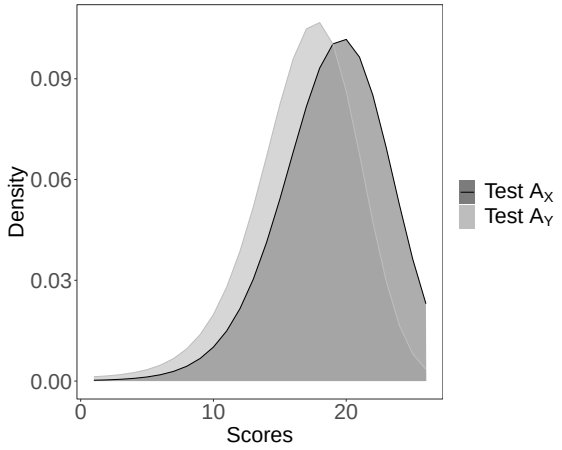
Figure 5: True total-score distributions for three distributional scenarios: (a) Symmetric, (b) Skewed, and (c) Bimodal.

of IRT model (Rasch, 2-PL, or 3-PL) is determined by the AIC, BIC and LRT, respectively, resulting in three potentially different parametrisations. Using these IRT models, the test forms are equated and the results are stored for each iteration.

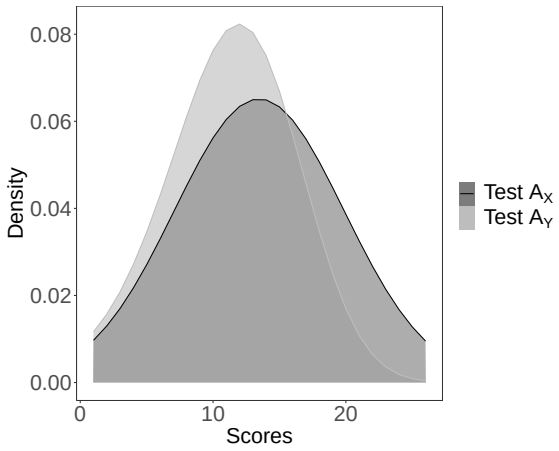
Since the aim of this paper is to inform practitioners on model selection for smoothing of test score distributions, and since IRT equating with kernel functions still lacks full implementation in R, we employ the traditional percentile-rank with the PSE method in this set-up. We note that this is a special case of kernel equating, using a uniform kernel with a fixed bandwidth of 0.33 (von Davier et al.; 2004). This method is available in several



(a) The anchor score probabilities generated from symmetric distributions.



(b) The anchor score probabilities generated from skewed distributions.



(c) The anchor score probabilities generated from bimodal distributions.

Figure 6: True distributions on anchor-set scores for three distributional scenarios: (a) Symmetric, (b) Skewed, and (c) Bimodal.

R packages for all IRT models that we consider in this paper. For this study, we use the **equateIRT** package (Battauz; 2015).

6.3 *Evaluation Measures*

The equating estimators are evaluated in terms of bias, simulation standard errors (SE) and the ASEE based on the asymptotic distribution of the equating function. Let $\hat{\varphi}(x)^{(r)}$ denote the kernel equating estimator evaluated at point x for the rth replicate using sample data,

let $\varphi(x)$ denote the population equating function and let

$$\bar{\varphi}(x) = \frac{1}{100}\sum_{r=1}^{100}\hat{\varphi}(x)^{(r)},$$

then

$$\text{Bias}\big[\hat{\varphi}(x)\big] = \frac{1}{100}\sum_{r=1}^{100}\hat{\varphi}(x)^{(r)} - \varphi(x), \tag{10}$$

and

$$\text{SE}\big[\hat{\varphi}(x)\big] = \sqrt{\frac{1}{100-1}\sum_{r=1}^{100}\big[\hat{\varphi}(x)^{(r)} - \bar{\varphi}(x)\big]^2}. \tag{11}$$

6.4 *Results – Set-Up A*

Figure 7 presents the percentage of accurately selected models ((X, A)-model, (Y, A)-model) across various model-selecting strategies for a sample size of $N = 3,000$. It provides insights into the effectiveness of each strategy across the three different distributional scenarios. For all distributions, the BIC and the ASEEmin criteria was most successful in model selection. For example, under the symmetric design, the BIC correctly identified 100% ((X, A)-model) and 96% ((Y, A)-model) of the models for $J = 40$ and 100% of the models for $J = 80$. The ASEEmin selector had almost the same accuracy. It is noteworthy that LRT consistently failed to correctly identify any model across all the categories and distributions. In summary, these results suggest that model selection performance is highly dependent on the choice of selection criterion, but not on the characteristics of the test score distributions and the test length. While BIC and ASEEmin appear most reliable, the AIC also shows high accuracy in selecting the true model. Whenever it fails to do so, it generally selects a bigger model with more cross-moments.

In Figures 8, 9, and 10, the percentages of times that the AIC, BIC and LRT, respectively, were chosen by the ASEEmin criterion are presented. In the symmetric data setting for the $N = 3,000$ case, the AIC selector appears to be used slightly more frequently than the BIC and LRT selectors in both the (X, A)-model and the (Y, A)-model. For the skewed data setting, the difference is clearer in favor of the AIC. For the bimodal data, the difference

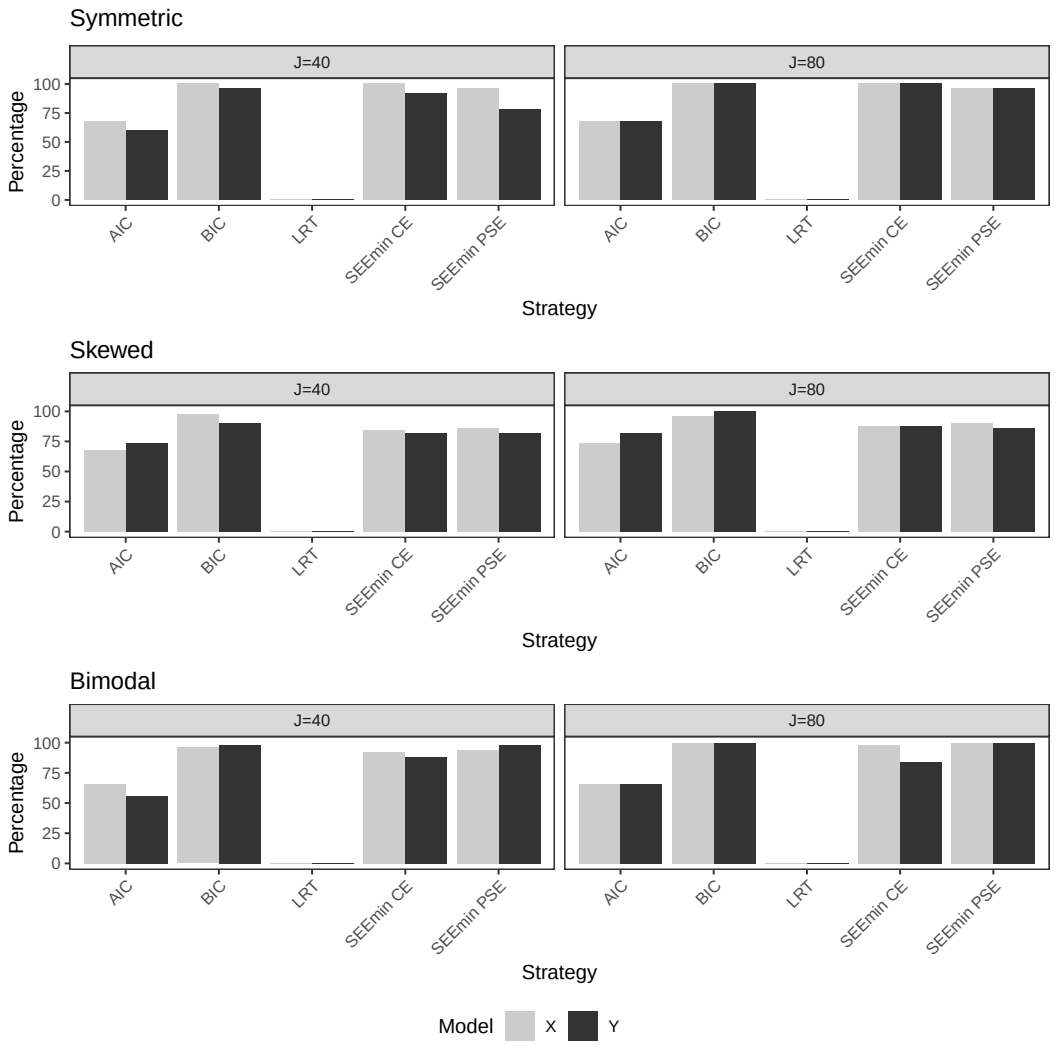


Figure 7: The percentage of correctly selected models ((X, A)-model, (Y, A)-model) for each respective model-selecting strategy for $N = 3,000$.

is rather small again, and with a few cases where the BIC is chosen more frequently. The LRT is, with only a few exceptions for the skewed data setting, never chosen, which clearly indicates that the ASEEmin favors model criteria with high model selection accuracy.

In Figure 11, the bias, SE and ASEE are displayed under the symmetrical distribution setting considering both the CE estimator (left panel) and the PSE estimator (right panel). As was seen in the Empirical Analysis, the estimators perform similarly along a majority of the score scale, and exhibit differences in the tails. In the upper range, the biases are the largest. The BIC and ASEEmin criteria show close to identical performance, especially for

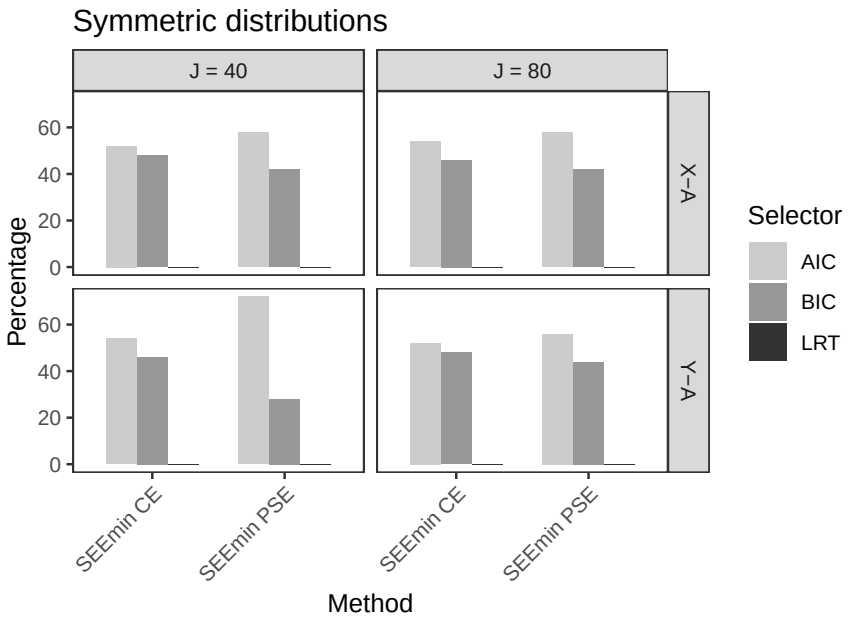


Figure 8: The percentage of times each selector was used in the proposed methods (X–A model, Y–A model) under the symmetric data setting, for $N = 3000$.

the CE estimator, since they selected very similar presmoothing models. Note that the AIC and BIC often select the same model, and when different, they select very similar models. The ASEEmin selects the AIC slightly more often for the symmetric design (see Figure 8) but the performance of the resulting equating estimator is very similar to both the AIC and the BIC-selected estimators. This is indeed reflected in Figure 11. Also note that the BIC has an almost perfect model selection performance in all of the distributional scenarios, so whenever the ASEEmin selects the BIC, which it does 40-50% of the time, it typically selects the true model. It will therefore be reflected in the results. Interestingly, the LRT has smaller bias than the AIC for the top scores, even though it consistently failed to select the true model. Instead, it seems like the slightly larger presmoothing models are not a disadvantage in terms of equating error. On the other hand, the LRT does produce the largest SE and ASEE values. We lastly notice that there is no clear winner between the CE and PSE estimator, as they both show similar performance.

As seen in Figures 12 and 13 the relative performance of the model-selection criteria remains similar for the skewed and bimodal data settings, where the sparse data at certain

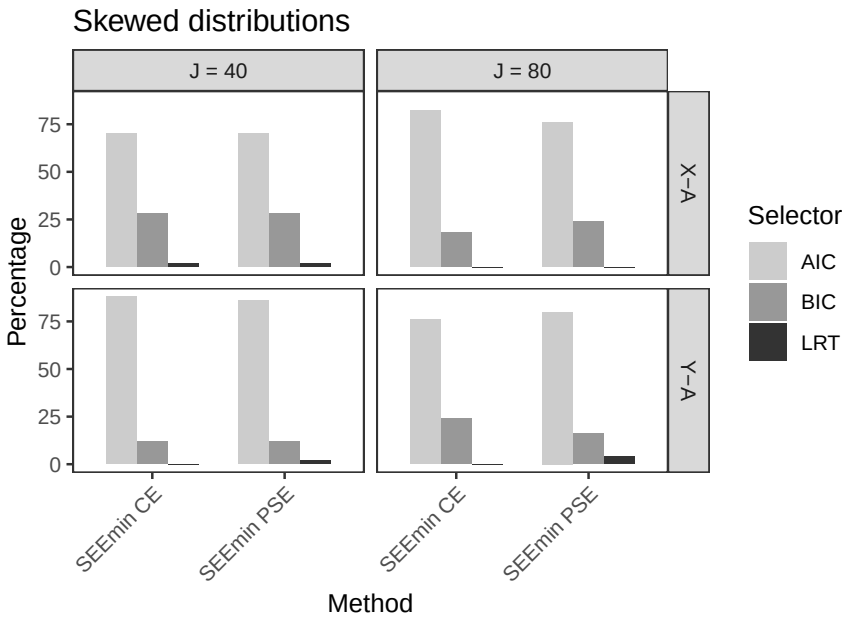


Figure 9: The percentage of times each selector was used in the proposed methods (X–A model, Y–A model) under the skewed data setting, for $N = 3,000$.

intervals from the respective distributions are reflected in increased equating error and uncertainty. It is noteworthy that the SE and ASEE for the equatings based on LRT-selected modes are particularly high in the tails of the score distributions for the skewed and bimodal settings.

6.5 *Results – Set-Up B*

In Figure 14, the percentage of correctly selected models in the IRT setting is displayed. Under the 1-PL data setting, the AIC and LRT criteria showed a distinct advantage in model selection, achieving 100% accuracy across all tested sample sizes and test lengths. In contrast, the BIC strategy demonstrated variable performance, with the percentage of correctly selected X–A and Y–A models fluctuating around 50-60%.

The 2-PL data setting saw all criteria correctly identifying the true model 100% of the times. The 3-PL data setting presented more challenging conditions for all three strategies, who performed the same model selection accuracy for both test lengths. The accuracy levels ranged from 10% to 46% depending on the test length, thus showing the overall worst

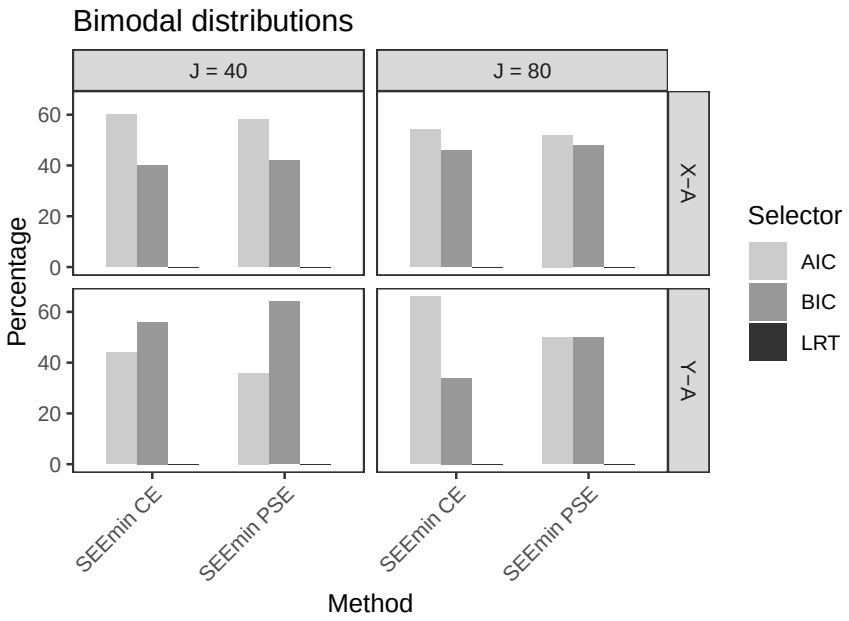


Figure 10: The percentage of times each selector was used in the proposed methods ((X, A)-model, (Y, A)-model) under the bimodal data setting, for $N = 3000$.

performance of all IRT settings. When the selection criteria selected an incorrect model in the 3-PL data generating case, they always selected the 2-PL model. We made sure to generate a guessing parameter that did not have values too close to 0, and so the 3-PL generated data would be different from the 2-PL generated data. It is however a challenging task to decide between the 2-PL and the 3-PL, a problem nicely discussed and investigated in Brown et al. (2015), and also reflected in our results. In summary, the AIC and LRT criteria stood out in the simpler 1-PL and 2-PL data settings, but all strategies experienced diminished performance in the more complex 3-PL setting.

In Figure 15, the bias, SE and ASEE are displayed under all three data-generating scenarios. We see that the bias of all equating estimators when the true data-generating process is either described by a 1-PL model or a 2-PL model is negligible. When the data is generated from a 3-PL model on the other hand, the bias increases. We also see that the simulation SEs and ASEE values generally increase as the complexity of the data-generating process increases. We conclude that when there is any practical difference between the methods, the LRT seem to perform the worst, followed by the AIC, and the BIC consistently performing

the best.

Finally, we note that the differences between the equated scores from the considered estimators were within the DTM for all considered scenarios in both simulation set-ups, which differs from our empirical analysis. We, however, note that the SE and ASEE values have clearer differences which could be of practical importance, especially in cases where the sample size is small.

7 Discussion

This study aimed to evaluate the impact of the AIC, BIC and LRT criteria for log-linear and IRT models when they are used to estimate the score distributions within the kernel equating framework. An algorithm is proposed which uses these three criteria and selects the pair which minimizes the ASEE. Since test groups are often heterogeneous, both within the group and between groups it is not necessarily the case that the same criterion will select the model which minimizes the estimated ASEE for both groups. We believe that the proposed criterion can serve as a tool to inform the user on which criterion that actually performs best in some well-defined sense. It could for example, in the best of cases, confirm that the selected model is the one that minimize the ASEE, or at least that the selected model is very similar to the one that minimize the ASEE. If several criteria point in the same direction, it might give further evidence in favor of a certain model. So even though it might be slightly unnatural to consider different criteria for different groups, we solve the model selection problem in an unconstrained way and use the ASEEmin criterion as a way to further inform us on the model selection. The study has considered both empirical and simulated data for the NEAT design and was motivated by the fact that the three model fit indices considered are all commonly used to select parameterization for these classes of models (Andersson and Wiberg; 2017; Moses and Holland; 2010a; Lêoncio et al.; 2023). Our findings, based on both real and simulated data, reveal that the choice of presmoothing model and model fit index impacts the equated scores, especially in terms of SE and ASEE, emphasizing their practical

importance.

Our analysis of log-linear models showed that model selection performance is dependent on the selection criterion. We discovered that the BIC and ASEEmin criteria were most effective in accurately selecting models across diverse distributional scenarios, which included symmetric, skewed, and bimodal data. While the AIC also demonstrated high accuracy in selecting the true model, it usually favored larger models with more cross-moments when it failed to do so. We, however, conclude that for both symmetric and skewed data, the AIC is the criterion which most often selects the pair of models which minimize the ASEE in finite samples. For bimodal data, the BIC instead most often minimize the ASEE.

Another intriguing finding was the relatively smaller bias displayed by LRT for top scores, despite its consistent failure in model selection. The LRT always selected a larger model than the true model, which implies that a larger presmoothing model may not inherently disadvantage equating error. However, caution is warranted as the LRT produced the highest SE and ASEE values, signaling potential inconsistencies in equating precision. We note that our implemented procedure is a slight alteration of the complex-to-simple strategy used in Moses and Holland (2009), where model selection of a univariate model, i.e., a log-linear model for X and Y, respectively, are considered. In Moses and Holland (2010a), bivariate data is considered; however, they fix the number of univariate moments and focus on model selection of the cross-moments. On the other hand, in our study, we consider model selection of the full model and implement an exhaustive search through all possible nested models. After all models are fitted and compared, our function iterates through the LRT results, comparing each p-value against the adjusted significance threshold as described in Section 4. The best model by LRT is determined as the one with a p-value below this threshold. If multiple models meet this criterion, the last one iterated over (and thus with the highest powers within the specified range) is selected. Since we tackle slightly different model selection problems (selecting the cross-moments vs. selecting the full model), the results are expected to differ from those of Moses and Holland (2009) and Moses and Holland (2010a).

In the IRT setting, the AIC and LRT criteria excelled in model selection in simpler 1-PL

and 2-PL data settings, achieving 100% accuracy. However, under the more complex 3-PL data scenario, all strategies witnessed reduced performance, with accuracy levels ranging from a mere 10% to 46%. This observation underscores the increasing challenge faced by these criteria when dealing with more complicated IRT models, and calls for further investigation into strategies that can maintain high model selection accuracy in these situations.

Concerning the bias, SE and ASEE, the BIC criterion consistently demonstrated the best performance in the IRT context. A discernible pattern emerged where the bias, as well as the SE and ASEE values, generally increased with the complexity of the data-generating process.

The idea of targeting the equating function when selecting log-linear presmoothing model has been considered recently by Liu and Kolen (2020), where a log-linear model-selection criterion aiming at minimizing an estimate of the mean squared error of the equating function was proposed. However, their method was only considered for univariate data under the EG design, which in general is a simpler task. Secondly, they aimed at minimizing a measure which requires the estimation of the equating bias, a quantity always unknown. The statistical properties of their selection method is therefore hard to determine. With the procedure proposed in Algorithm 1 there are certain statistical guarantees since we only consider models among the subset of models that have been selected by the AIC, BIC and LRT in the first step, which all have good and theoretically established selection properties.

In conclusion, our study emphasizes the important role of model selection in test score equating and highlights the importance of the careful choice of model fit indices. Given the variable performance of different criteria under different distributional scenarios and test designs, a one-size-fits-all strategy might be inadvisable. We therefore recommend that practitioners test the sensitivity of their equating results to slight changes in the presmoothing model. One convenient way of doing so is to use our own R function for log-linear model-selection, which takes the bivariate score data as input and outputs the best fitting model according to the AIC, BIC, LRT and ASEEmin, for any number of univariate and bivariate moments, as specified by the user. We note that the ASEE is merely one of several possible

equating-specific measures that can be used in the proposed algorithm. In our R function, we also give the user the possibility to select the model which minimizes the percent relative error (von Davier et al.; 2004), another equating-specific evaluation measure. In our simulations, the results of such criterion did however not stand out but performed similarly to the other criteria, and are therefore omitted. Further research on this and similar criteria is however motivated. It would also be of interest to examine which measure to use when including covariates such as age and gender in the presmoothing models, as described in Wiberg and Bränberg (2015) and Wallin and Wiberg (2019). In these cases the models may be more complicated, or misspecified as in Wallin and Wiberg (2023) and thus more research is needed. One limitation to our study is that we have only considered external anchor items, motivated by the design of the empirical data. By considering internal anchor items, one has to, additionally, address the issue of structural zeros. Model-selection performance under such setting is left for future research. Finally, we point out that this study has not considered postsmoothing, i.e., smoothing of the equipercentile transformation rather than the test score probabilities. The topic of postsmoothing is left for future research.

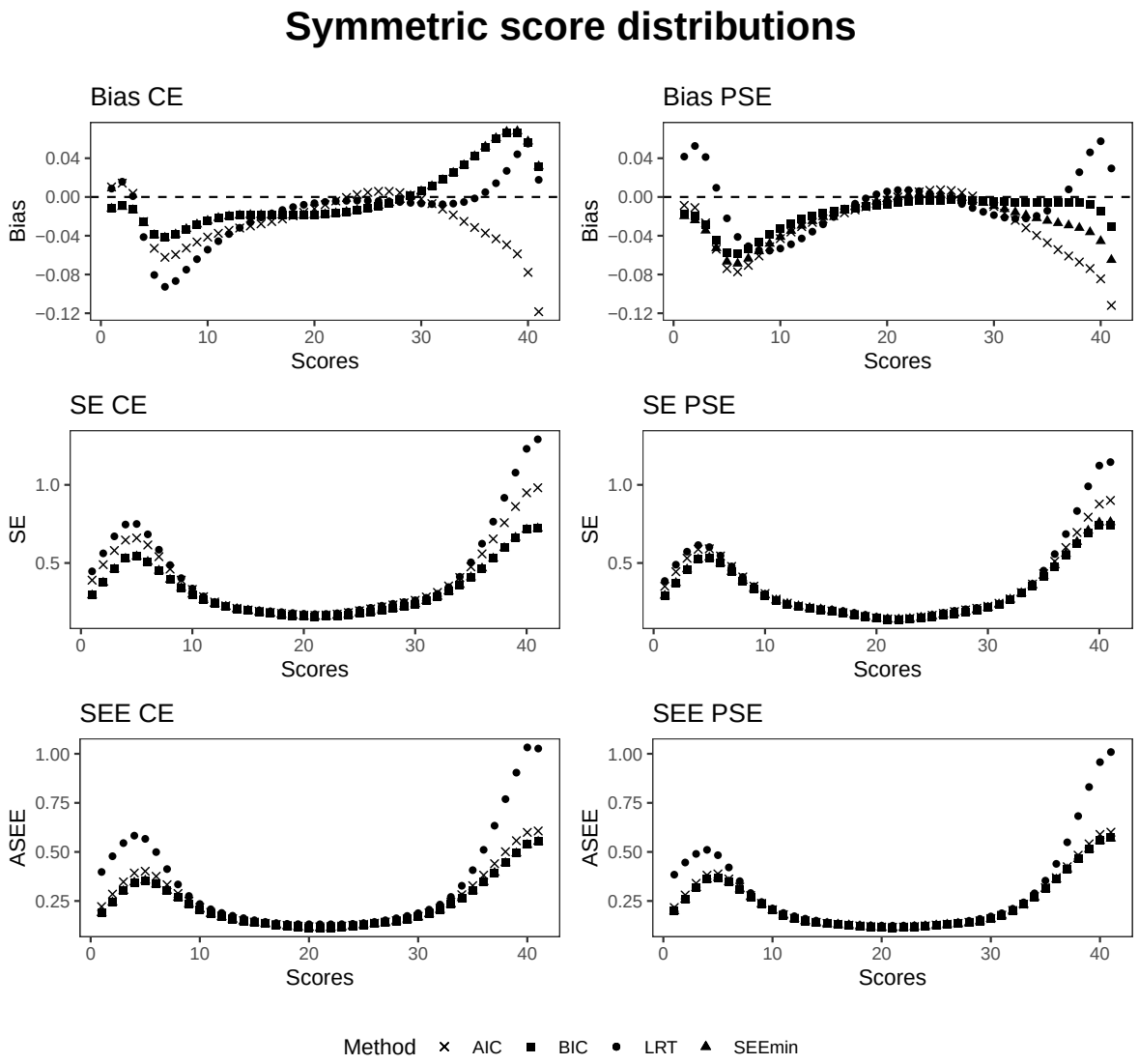


Figure 11: The bias, SE and ASEE of each equating estimator for a test length of 40, a sample size of 3,000 and symmetric score probability distributions.

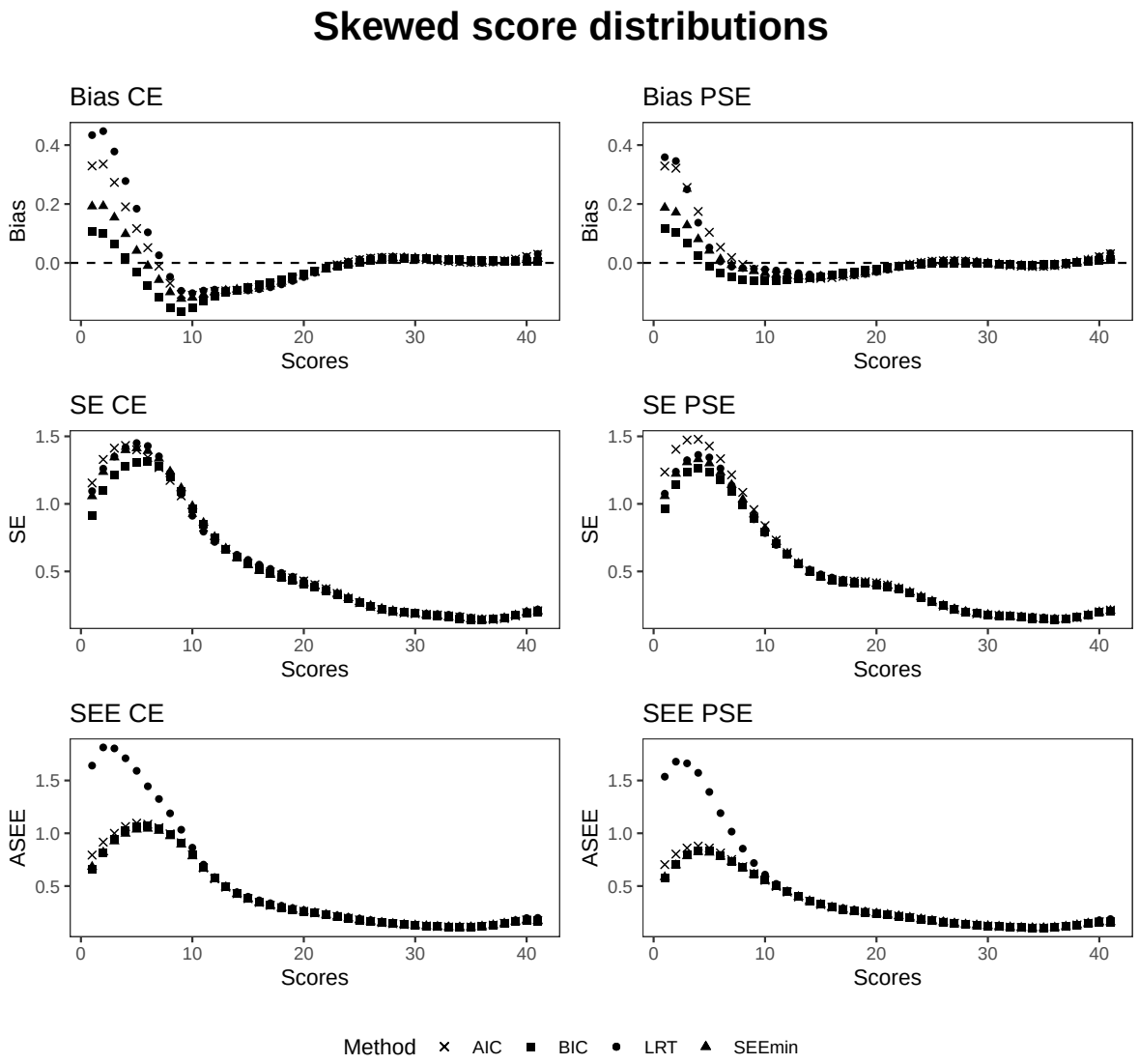


Figure 12: The bias, SE and ASEE of each equating estimator for a test length of 40, a sample size of 3,000 and skewed score probability distributions.

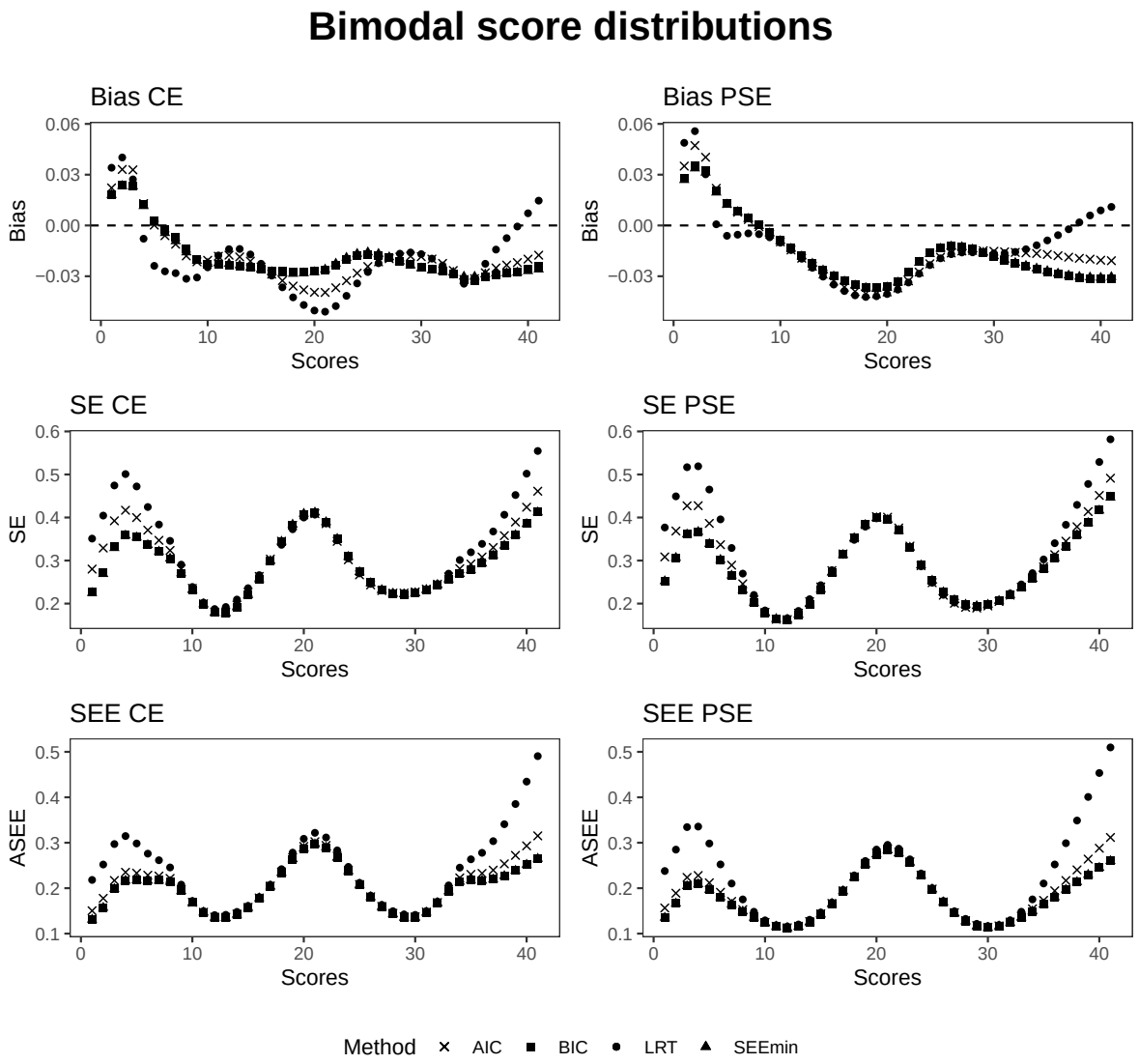


Figure 13: The bias, SE and ASEE of each equating estimator for a test length of 40, a sample size of 3,000 and bimodal score probability distributions.

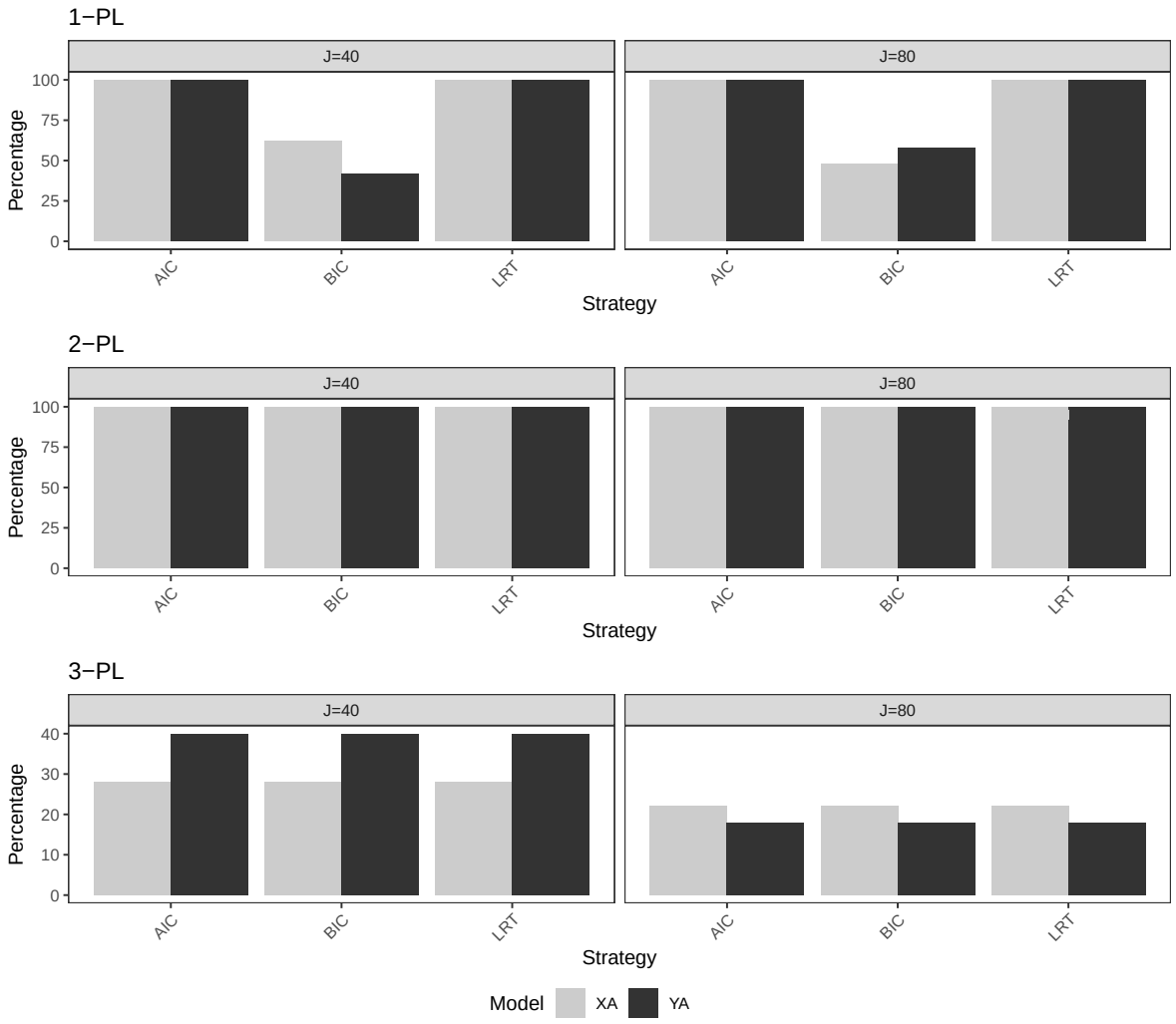


Figure 14: The percentage of correctly selected models ((X, A)-model, (Y, A)-model) for each respective model-selecting strategy (in the IRT setting) for $N = 3000$.

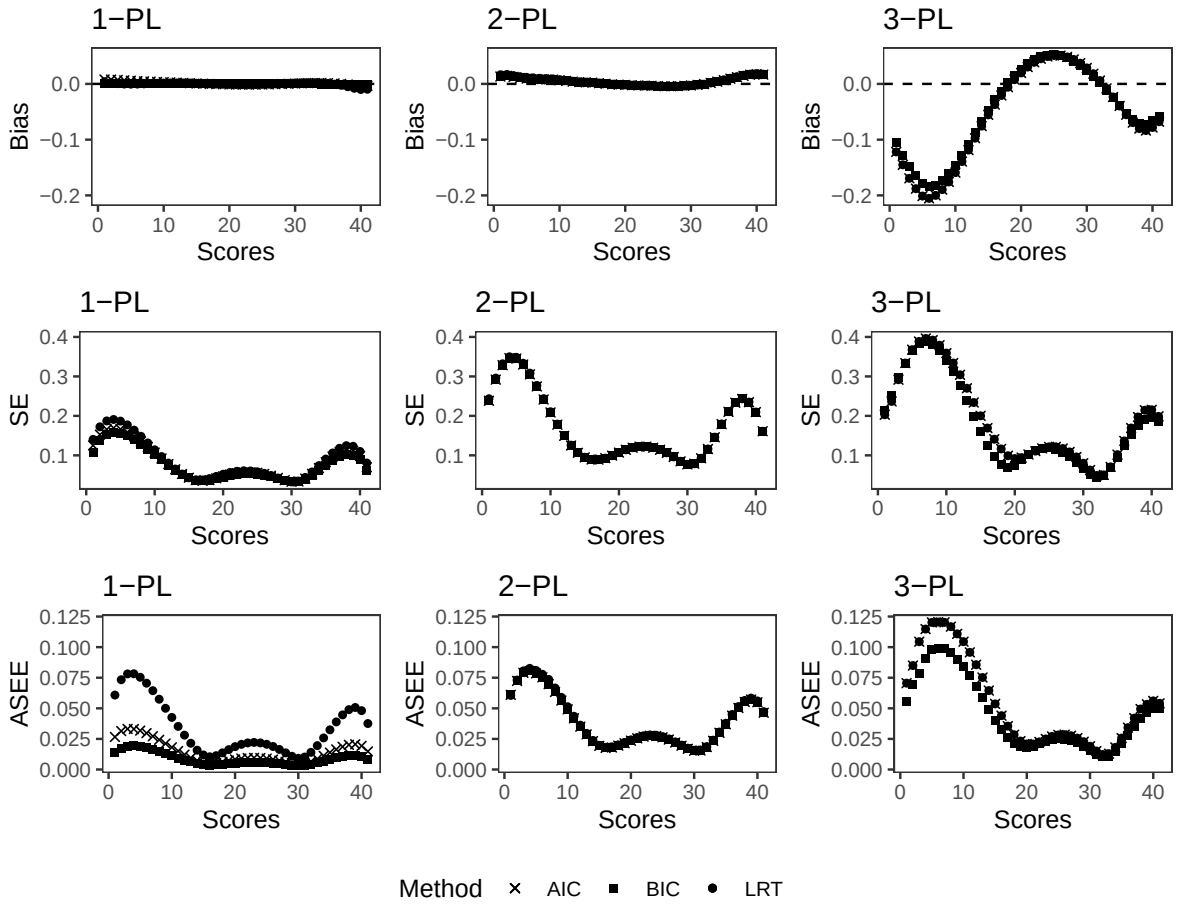


Figure 15: The bias, SE and ASEE under the 1-PL, 2-PL and 3-PL data-generating setting, for $N = 3000$ and $J = 40$.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

Andersson, B., Bränberg, K. and Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate, *Journal of Statistical Software* **55**(6): 1–25.

Andersson, B. and Wiberg, M. (2017). Item response theory observed-score kernel equating, *Psychometrika* **82**(1): 48–66.

Battauz, M. (2015). equateirt: An r package for irt test equating, *Journal of Statistical Software* **68**: 1–22.

Birnbaum, A. (1968). Some latent trait models and their use in inferring any examinee's ability, *in* F. M. Lord and M. R. Novick (eds), *Statistical theories of mental test scores*, Reading, MA: Adison-Wesley, pp. 395–479.

Brown, C., Templin, J. and Cohen, A. (2015). Comparing the two- and three-parameter logistic models via likelihood ratio tests: A commonly misunderstood problem, *Applied Psychological Measurement* **39**(5): 335–348.

Cui, Z. and Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method, *Journal of Educational Measurement* **46**(2).

Dorans, N. and Feigenbaum, M. (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. and Wiberg, M. (2017). *Applying test equating methods using R*, New York: Springer.

González, J., Wiberg, M. and von Davier, A. A. (2016). A note on the Poisson's binomial distribution in item response theory, *Applied Psychological Measurement* **40**(4): 302–310.

Haberman, S. J. (1974a). *The analysis of frequency data*, University of Chicago Press.

Haberman, S. J. (1974b). Log-linear models for frequency tables with ordered classifications, *Biometrics* **30**(4): 589–600.

Hambleton, R. K. and Swaminathan, H. (1985). *Item response theory: Principles and applications*, Dordrecht: Kluwer Nijhoff Publishing.

Hanson, B. A. (1991). A comparison on bivariate smoothing methods in common-item equipercentile equating, *Applied Psychological Measurement* **15**(4): 391–408.

Holland, P. and Thayer, D. (1987). Notes on the use of log-linear models for fitting discrete probability distributions, *ETS Research Report Series* **1987**(2): i–40.

Holland, P. and Thayer, D. (1989). The kernel method of equating score distributions, *Technical report*, Princeton, NJ: Educational Testing Service.

Holland, P. and Thayer, D. (2000). Univariate and bivariate loglinear models for discrete test score distributions, *Journal of Educational and Behavioral Statistics* **25**(2): 133–183.

Kim, D.-I., Brennan, R. and Kolen, M. (2005). A comparison of irt equating and beta 4 equating, *Journal of Educational Measurement* **42**(1): 77–99.

Kolen, M. (1991). Smoothing methods for estimating test score distributions, *Journal of Educational Measurement* **28**(3): 257–282.

Kolen, M. and Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices*, 3rd edn, New York: Springer.

Lêoncio, W., Wiberg, M. and Battauz, M. (2023). Evaluating equating transformations in irt observed-score and kernel equating methods, *Applied psychological measurement* **47**(2): 123–140.

Liu, C. and Kolen, M. J. (2020). A new statistic for selecting the smoothing parameter for polynomial loglinear equating under the random groups design, *Journal of Educational Measurement* **3**(27): 458–479.

Livingston, S. (1993). Small-sample equatings with log-linear smoothing, *Journal of Educational Measurement* **30**(1): 23–39.

Lord, F. and Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings", *Applied Psychological Measurement* **8**(4): 453–461.

Lyrén, P.-E. and Hambleton, R. K. (2011). Consequences of violated equating assumptions under the equivalent groups design, *International Journal of Testing* **11**(4): 308–323.

Moses, T. and Holland, P. W. (2007). Kernel and traditional equipercentile equating with degrees of presmoothing, *(Research Report No. RR-07-15). Princeton, NJ: Educational Testing Service* .

Moses, T. and Holland, P. W. (2009). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy, *Journal of Educational Measurement* **46**(2): 159–176.

Moses, T. and Holland, P. W. (2010a). A comparison of statistical selection strategies for univariate and bivariate log-linear models, *British Journal of Mathematical and Statistical Psychology* **63**(3): 557–574.

Moses, T. and Holland, P. W. (2010b). The effects of selection strategies for bivariate log-linear smoothing models on neat equating functions, *Journal of Educational Measurement* **47**(1): 76–91.

Moses, T. and Liu, J. (2011). Smoothing and equating methods applied to different types of test score distributions and evaluated with respect to multiple equating criteria, *(Research Report No. RR-11-20). Princeton, NJ: Educational Testing Service.* .

R Core Team (2022). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Read, T. R. and Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*, Springer-Verlag New York.

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

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**: 461–464.

von Davier, A., Holland, P. and Thayer, D. (2004). *The kernel method of test equating*, New York: Springer.

Wallin, G., Häggström, J. and Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating?, *Applied Psychological Measurement* **45**(7–8): 518–535.

Wallin, G. and Wiberg, M. (2019). Propensity scores in kernel equating under the non-equivalent groups with covariates design, *Journal of Educational and Behavioral Statistics* **44**(4): 390–414.

Wallin, G. and Wiberg, M. (2023). Model misspecification and robustness of observed-score test equating using propensity scores, *Journal of Educational and Behavioral Statistics* **48**(5): 603–635.

Wiberg, M. and Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design, *Applied Psychological Measurement* **39**(5): 349–361.

Wilk, M. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data, *Biometrika* **55**(1): 1–17.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula, *Journal of Statistical Software* **21**: 1–21.