Combining Propensity Scores and Common Items for Test Score Equating

Abstract

Ensuring that test scores are fair and comparable across different test forms and different test groups is a significant statistical challenge in educational testing. Methods to achieve score comparability, a process known as test score equating, often rely on including common test items or assuming that test taker groups are similar in key characteristics. This study explores a novel approach that combines propensity scores, based on test takers' background covariates, with information from common items using kernel smoothing techniques for binary-scored test items. An empirical analysis using data from a high-stakes college admissions test evaluates the standard errors and differences in adjusted test scores. A simulation study examines the impact of factors such as the number of test takers, the number of common items, and the correlation between covariates and test scores on the method's performance. The findings demonstrate that integrating propensity scores with common item information reduces standard errors and bias more effectively than using either source alone. This suggests that balancing the groups on the test-takers' covariates enhance the fairness and accuracy of test score comparisons across different groups. The proposed method highlights the benefits of considering all the collected data to improve score comparability.

Keywords: Educational testing, academic admission, fairness, equating, NEAT design

Introduction

Assessment tests in education are important tools for measuring students' knowledge, skills, and development. These tests also play a significant role in educational decision-making, influencing everything from teaching practices to college admissions. Given their impact, it is essential to ensure that test score interpretations are both valid and fair (see Chapter 3, American Educational Research Association et al., 2014). When test forms change or when different groups take different test forms, ensuring fair and comparable scores becomes a significant statistical challenge.

To address this challenge, traditional methods for adjusting scores often rely on including common items in the tests - known as anchor items - or assuming that the groups being compared are similar in their distributions of the latent ability the assessment test is designed to measure. These methods aim to adjust for variations in test difficulty and differences in group

abilities. However, when no common items are available, these methods may not fully account for group differences, potentially having significant impacts on, for example, academic admission decisions. Complicating matters further, the latent trait levels of the test-takers are not directly observable, which makes it non-trivial to condition the analysis on their values.

In light of these challenges, test score equating has emerged as a routine statistical process for most large-scale testing programs around the world. Test equating methods, used to align scores from different test forms onto a common scale, account for variations in test difficulty and differences in the ability levels of test-taking groups (González & Wiberg, 2017). The choice of equating method depends on assumptions about the test-takers and the available data. When groups of test takers receiving different test forms can be assumed to be similar in their distributions of the ability the test is designed to measure, the Equivalent Groups (EG) design can be used. However, if these groups cannot be assumed equivalent but have completed a set of common items (an anchor test), the nonequivalent groups with anchor test (NEAT) design is suitable (von Davier, Holland, and Thayer, 2004). When test-taking groups are not similar and no anchor test is administered, but information about test takers' covariates is available, the nonequivalent groups with covariates (NEC) design can be employed (Wiberg & Bränberg, 2015). Examples of tests with non-equivalent groups but without anchor items include the Invalsi test (INVALSI, 2013), the Armed Services Vocational Aptitude Battery (Quenette et al., 2006), and, until 2011, the Swedish Scholastic Aptitude Test (SweSAT; Stage & Ögren, 2004).

The importance of flexible equating methods became even more apparent during the global spread of Covid-19, which created unprecedented challenges for many large-scale assessments. For instance, the SweSAT faced restrictions on test-taker eligibility, resulting in new demographics taking the test (Wiberg, Lyrén & Lind Pantzare, 2021). Despite these changes in the test-taking population, the need to compare scores with previous administrations remained crucial, given the role of SweSAT in college admissions. Historically, researchers addressing changing background distributions of test groups have employed either the NEAT design when an anchor test was available, or the NEC design. However, only a few attempts have been made to integrate information from both covariates and anchor tests, highlighting a gap in the current methodology. Notable exceptions include Wiberg and Bränberg (2015), who explored a case of merging NEAT and NEC designs using categorical covariates and anchor test scores, and Albano and Wiberg (2019), who examined traditional equating methods combining anchor scores with a single covariate. Further, Lu and Guo (2018) used simulations to include information from an anchor test with pseudo equivalent groups (PEG) in a NEAT design. They

concluded that if the ability group difference were large, to use only NEAT design was to be preferred over PEG, but the NEAT linking could be improved by using PEG procedures based on background variables. If group ability differences were small, PEG linking produced comparable results to NEAT. Further, Lu and Kim (2021) used statistically matching equating samples in a NEAT design, while Kim and Walker (2021) used PEG when examining nonequivalent groups caused by suboptimal randomization, a short anchor test of only five items, and minimal collateral information. Recently, Kim and Walker (2022) extended this study when building on the PEG approach and used resampling to evaluate the linking accuracy of group adjustment using sample weights via minimum discriminant information adjustment (MDIA) using test takers' demographic information, a three-item anchor test, and a mixture of both. They concluded that using both sample weights via MDIA and a short anchor produced the most accurate equating results. More recently, Ozsoy and Kilmen (2023) compared NEAT and NEC designs in a modern equating framework, however they did not combine the two designs.

A promising approach to incorporating information about test takers from covariates is to use propensity scores. Define for each test taker the propensity score $e(\mathbf{D})$, which is the probability of being assigned a specific treatment (in this case test form) given the covariate vector $\mathbf{D}$ (Rosenbaum & Rubin, 1983). Set a treatment variable Z equal to 1 if test form Y (active treatment) is administered, and equal to 0 if test form X (control treatment) is administered. Then, the propensity score is defined as $e(\mathbf{D})=\Pr(Z = 1|\mathbf{D})$. If $\mathbf{D}$ contains every confounder of the relationship between (X, Y) and Z, the propensity score is a balancing score and it is enough to control for $e(\mathbf{D})$ to create balance in the test groups. The first study to consider propensity scores in test equating was Livingston, Dorans, and Wright (1990), who used them for sample matching. This approach was further developed by Yu, Livingston, Larkin, and Bonett (2004) and Paek, Liu, and Oh (2006). Subsequent researchers expanded these proposals, with Sungworn (2009) and Powers (2010) using propensity scores to improve traditional equating methods. Moses, Deng, and Zhang (2010) took a different approach, using propensity scores to combine two anchor test scores rather than incorporating external covariates in the analysis. Longford (2015) proposed equating based on matching with either inverse proportional weighting or matched pairs, derived from propensity scores based on background variables, while Haberman (2015) employed propensity scores to create PEG from nonequivalent groups before conducting equating.

Wallin and Wiberg (2019) proposed to use propensity scores with the NEAT design, framed within a modern equating framework building on kernel smoothing techniques. Their work

demonstrated that stratifying on the propensity scores, an idea dating back to Rosenbaum and Rubin (1984), could achieve a similar level of precision and accuracy compared to the NEAT design, provided the propensity scores are known. Recognizing that propensity scores are never truly known in practice, Wallin and Wiberg (2023) conducted a sensitivity analysis of equated scores to various misspecifications in the propensity score model. Their findings revealed that omitting an important covariate leads to biased estimates of the equated scores, while misspecifying a nonlinear relationship between covariates and test scores increases the equating standard error in the tails of the score distributions. Encouragingly, they also found that the equating estimators are robust against omitting a second-order term and using an incorrect link function in the propensity score estimation model.

Building upon this rich body of research, our paper introduces a novel approach in test equating by combining propensity scores with anchor test scores within the generalized kernel equating framework (Wiberg, González, & von Davier, 2025). An important reason to use kernel equating here is that kernel equating methods are used in practice to equate the college admissions test which we use in the empirical study. While recent studies have utilized propensity scores in kernel equating, none have explored the integration of both propensity scores and anchor test scores in this context, as proposed here. Our overall aim is to examine kernel equating with binary scored items when using propensity scores together with anchor test scores and covariates, comparing this approach with using either only anchor scores in the NEAT design or only propensity scores with the NEC design. We conduct both an empirical study and a simulation study. This allows us to assess the practical implications of our method in a real-world context while also investigating the bias, root mean squared error, and standard errors under varying conditions.

The rest of this paper is structured as follows. In the next section, kernel equating in general is described, followed by a description of kernel equating with propensity scores. This is followed by an empirical study with some results and a simulation study. The last section contains a discussion with some concluding remarks and practical implications.

Kernel equating

Kernel equating (von Davier, Holland & Thayer, 2004; Wiberg, et al., 2025) aims to equate test score X to test score Y on a target population T. For the NEAT and the NEC design, the target population T is not trivial to define since we are dealing with samples from two distinct population, P and Q. It is common to define a synthetic target population, defined symbolically

as $T = wP + (1 - w)Q$, with $0 \leq w \leq 1$. In practice, $w > 0$ is typically used to ensure comparability across administrations.

Kernel equating comprises five steps: 1) Presmoothing, 2) Estimation of the score probabilities, 3) Continuization, 4) Equating, and 5) Evaluating the equating transformation. Denote the observations of *X* and *Y* by x_j, $j = 1, \ldots, J$, and y_k, $k = 1, \ldots, K$, respectively. Let $r_j = \Pr(X = x_j|T)$ and $s_k = \Pr(Y = y_k|T)$ be the probabilities of a randomly selected test-taker in the target population *T* scoring x_j on test form X and y_k on test form Y, respectively. In the first presmoothing step, a log-linear model is typically fitted to the data to reduce the sampling variance. For the NEAT design, denote the observations of anchor test *A* by a_l, $l = 1, \ldots, L$, and define the joint probability as $p_{jl} = \Pr(X = x_j, A = a_l)$, then

$$\log(p_{jl}) = \beta_0 + \sum_{i=1}^{T_r} \beta_{x,i} x_j^i + \sum_{k=1}^{T_a} \beta_{x,i} x_j^i + \sum_{d}^{T_{xa}} \sum_{d'}^{T_{ax}} \beta_{xa,dd'} x_j^d a_l^{d'} \quad (1)$$

By estimating the parameters using maximum likelihood estimation, the sample moments are preserved in the distribution being modelled. Several models are typically fitted and the best fitting model according to some criteria is chosen. From the fitted model we obtain the estimated test score probabilities in step 2. If some other proxy of ability is available, such as a propensity score, these can also be modelled in the presmoothing model which will be demonstrated later in the paper.

To obtain the equating transformation, which maps the test scores onto a common scale, we define the cumulative distribution functions (CDFs) of *X* and *Y* in *T* as $F(x) = \Pr(X \leq x|T)$ and $G(y) = \Pr(Y \leq y|T)$, respectively. Kernel equating defines equivalent scores as those that share the same relative position in their respective distributions, using the equipercentile equating transformation:

$$y = \varphi_Y(x) = {G_Y}^{-1}(F_X(x)). \quad (2)$$

The equipercentile transformation is the most commonly used method to equate test scores among large-scale testing organizations. As test scores are discrete, continuous approximations of the test score distributions are typically utilized. Kernel equating utilizes kernel functions for this purpose, most commonly a Gaussian kernel function. Let $\Phi(\cdot)$ represent the standard normal distribution function, and $\boldsymbol{r} = (r_1, \ldots, r_J)^t$, then the continuized CDF for score *X* is defined as

$$F_{h_X}(x; \boldsymbol{r}) = \Pr(X(h_X) \leq x) = \sum_j r_j \, \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right),$$

where, $\mu_X = \sum_j x_j r_j$ is the mean of X in population T, $a_X = \sqrt{\sigma_X^2/(\sigma_X^2 + h_X^2)}$, σ_X^2 is the variance of X in population T, and $h_X > 0$ is the bandwidth which determines the smoothness level of the continuous approximation. The bandwidth can be selected in several ways, and for a comparison of different bandwidth selection methods see Wallin, Häggström and Wiberg (2021). The continuization of the Y score distribution to obtain $G_{h_Y}(y; \boldsymbol{s})$, $\boldsymbol{s} = (s_1, \ldots, s_K)^t$, is done in an analogous way. Equation 2 is then used to carry out the equating with these continuized CDFs.

$$\hat{\varphi}_Y(x) = G_{h_Y}^{-1}(F_{h_X}(x)) \tag{3}$$

Finally, the equating transformation can be evaluated with different measures, including the asymptotic standard error of equating (SEE; von Davier, et. al., 2004), which, using the delta method, is defined as

$$\text{SEE}(x) = \sqrt{\text{Var}(\hat{\varphi}_Y(x))} = ||\mathbf{J}_{\varphi_Y}\mathbf{J}_{\text{DF}}\mathbf{C}||. \tag{4}$$

The term $\hat{\varphi}_Y(x)$ is defined in Equation 3, $\mathbf{J}_{\varphi_Y}$ represents the Jacobian matrix of the equating function, $\mathbf{J}_{\text{DF}}$ denotes the Jacobian matrix of the design function, and $\mathbf{C}$ is defined such that $\text{cov}(v(\boldsymbol{P}), v(\boldsymbol{Q})) = \boldsymbol{C}\boldsymbol{C}'$, where $\boldsymbol{P} = \{p_{jl}\}_{J\times L}$ and $\boldsymbol{Q} = \{q_{kl}\}_{K\times L}$ and $v(\cdot)$ denotes the vectorization of a matrix, where the columns are stacked on top of each other. The design function is defined such that $(\boldsymbol{r}, \boldsymbol{s})' = DF(\boldsymbol{P}, \boldsymbol{Q})$ and is, as the name suggests, design specific. See Wallin and Wiberg (2019) and von Davier et al. (2004) for the specific function specification for the NEAT design and NEC design with propensity scores. Lastly, note that the SEE definition gives us a standard error value for each test score x. The SEE therefore typically reflects the naturally occurring sparsity of data in the tails of the score distributions (only very few test-takers get a score of 0 or close to 0, and likewise for the highest scores).

Kernel equating with the NEAT design and categorized covariates in the NEC design

To perform kernel equating in the NEAT design we have two choices. First, we can utilize the mixture definition of the target population T to construct distributions of X and Y in T and obtain test score probabilities:

$$r_j = \Pr(X = x_j \mid T) = wr_{Pj} + (1-w)r_{Qj}, \tag{5}$$

and

$$s_k = \Pr(Y = y_k \mid T) = ws_{Pk} + (1-w)s_{Qk}, \tag{6}$$

where $r_{Pj} = \Pr(X = x_j \mid P)$, $r_{Qj} = \Pr(X = x_j \mid Q)$, $s_{Pk} = \Pr(Y = y_k \mid P)$ and $s_{Qk} = Pr(Y = y_k \mid Q)$ are the score probabilities of *X* and *Y* in populations *P* and *Q*, respectively. We can then equate the obtained distributions using kernel poststratification equating (KPSE) using Equation 3 directly. Secondly, we can link the different test forms through a chain and thus obtain kernel chained equating (KCE), defined as

$$\varphi_Y(x) = G_{h_Y}^{-1}(H_{h_Y}(H_{h_X}^{-1}(F_{h_X}(x)))) \quad (7)$$

where H_{h_Y} and H_{h_X} are the continuized CDFs for the anchor test forms given to the group that received test form X and test form Y.

If we are using categorized covariates, as in Wiberg and Bränberg (2015), we just exchange the anchor test scores in Equations 3 and 7 to the categorized covariate information. How to proceed if we instead of categorized covariates use propensity scores is described next.

Kernel equating with propensity scores

Wallin and Wiberg (2019) proposed the use of propensity scores in the NEC design with both KPSE and KCE estimators and further expanded the theory in Wallin and Wiberg (2023). One advantage of using propensity scores in test equating is that they summarize multiple covariates into a single scalar, thereby reducing the dimensionality of the problem. This is particularly important when incorporating background variables in log-linear smoothing models, as modelling each covariate directly can lead to sparsity issues - many combinations of test scores and covariate values may have few or no observations, making parameter estimation unstable. By using propensity scores, we avoid sparsity issues while still adjusting for observed confounders.

A fundamental property of propensity scores is that if $\boldsymbol{D}$ contains all confounders of the relationship between test form assignment and test scores then conditioning on $e(\boldsymbol{D})$ is sufficient to balance the groups. Specifically, we assume that:

$$P\left(X = x_j \middle| Z = 1, A, e(\boldsymbol{D})\right) = P\left(X = x_j \middle| Z = 0, A, e(\boldsymbol{D})\right),$$

which implies that once we control for the anchor score A and the propensity score $e(\boldsymbol{D})$, any remaining differences between the groups are random rather than systematic. This balancing assumption allows us to compare test scores fairly between groups, even when direct matching on all covariates is not feasible.

There are multiple ways to estimate propensity scores. In this paper, we use logistic regression, following the common approach of subdividing test takers into strata based on the percentiles of their estimated propensity scores (Rosenbaum & Rubin, 1984). Within each

stratum, test takers are assumed to be comparable in ability. The number of strata is chosen based on the covariate distribution to ensure adequate balancing while maintaining a sufficient number of observations in each stratum.

Other methods for balancing covariates include weighting techniques, such as the minimum-variance balancing method proposed by Zubizareta (2015), which adjusts the empirical distribution of covariates to achieve a prespecified level of balance. Additionally, a range of quantitative and qualitative diagnostics can be used to assess balance between test forms after weighting or stratification. For a comprehensive review of propensity score methods, we refer to Austin and Stuart (2015).

The KPSE estimator with propensity scores

To obtain a KPSE estimator with propensity scores, i.e., the PS-KPSE estimator, denote the stratified propensity score for strata l, $l = 1, \ldots, L$, by $e_{Xl}(\mathbf{D})$ and $e_{Yl}(\mathbf{D})$ for populations P and Q, respectively. Let $\mathbf{d}$ represent the observed value of $\mathbf{D}$ and let $p_{jl} = \Pr(X = x_j, e(\mathbf{D}_{Xl}) = e(\mathbf{d}_{Xl}) \mid P)$ and $q_{kl} = \Pr(Y = y_k, e(\mathbf{D}_{Yl}) = e(\mathbf{d}_{Yl}) \mid Q)$ denote the joint probabilities of the test scores and the categorized propensity scores for population P and Q, respectively. r_{Qj} and s_{Qk} can be estimated directly through $\hat{r}_{Pj} = \sum_l \hat{p}_{jl}$ and $\hat{s}_{Qk} = \sum_l \hat{q}_{kl}$. By design, there is no data to estimate r_{Qj} and s_{Pk} but if we assume that the conditional distributions of X given $e(\mathbf{D})$ and Y given $e(\mathbf{D})$ is the same in population P and Q respectively they can be estimated as follows

$$\hat{r}_{Qj} = \sum_l \left(\frac{\hat{p}_{jl}}{\sum_j \hat{p}_{jl}} \cdot \sum_k \hat{q}_{kl} \right) \text{ and } \hat{s}_{Pk} = \sum_l \left(\frac{\hat{q}_{kl}}{\sum_k \hat{q}_{kl}} \cdot \sum_j \hat{p}_{jl} \right). \quad (8)$$

Equations 8 are then plugged into Equations 3, 5 and 6 and we can obtain the PS-KPSE estimator as follows

$$\varphi_Y(x; \hat{\boldsymbol{r}}, \hat{\boldsymbol{s}})_{\text{PSE}} = G_{h_Y}^{-1}\left(F_{h_X}(x; \hat{\boldsymbol{r}}); \hat{\boldsymbol{s}}\right). \quad (9)$$

The CE estimator with propensity scores

To define an estimator when using CE with propensity scores, i.e. the PS-KCE estimator, define the continuized CDFs for X and Y in population Q as $F_{h_P}(x; \hat{\boldsymbol{r}}_P) = \hat{F}_{h_P}(x)$, and $G_{h_Q}\left(y; \hat{s}_Q\right) = \hat{G}_{h_Q}(y)$, where $\boldsymbol{r}_P = (r_{P1}, \ldots, r_{PJ})^t$ and $\boldsymbol{s}_Q = (s_{Q1}, \ldots, s_{QK})^t$. However, we also need to define the continuized CDFs H for the anchor tests $H_{h_{e_{Xl}}}(e_{Xl}(\boldsymbol{d}); \hat{t}_P) = \widehat{H}_{h_{eXl}}(e_{Xl}(\boldsymbol{d}))$, and $H_{h_{e_{Yl}}}\left(e_{Yl}(\boldsymbol{d}); \hat{\boldsymbol{s}}_Q\right) = \widehat{H}_{h_{eYl}}(e_{Yl}(\boldsymbol{d}))$, with score probabilities$t_P = (t_{P1}, \ldots, t_{PL})^t$ and $t_Q =$

$\left(t_{Q1},\ldots,t_{QL}\right)^t$, where $t_{Pl} = \Pr(e_{Xl}(\boldsymbol{D}) = e_{Xl}(\boldsymbol{d})|P)$ and $t_{Ql} = \Pr(e_{Yl}(\boldsymbol{D}) = e_{Yl}(\boldsymbol{d})|Q)$. The PS-KCE estimator can then be defined as

$$\hat{\varphi}_{Y(CE)}(x) = \varphi_{Y(CE)}\left(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{t}}_Q,\hat{\boldsymbol{s}}_Q\right) = \hat{G}^{-1}_{h_{YQ}}(\hat{H}_{h_{e_{Yl}}}(\hat{H}^{-1}_{h_{e_{Xl}}}(\hat{F}_{h_{XP}}(x)))). \quad (10)$$

Combining anchor test and covariate information

Anchor test information can be incorporated in at least two different ways when using kernel equating with propensity scores. Either they can be incorporated directly in the propensity score model (labelled PSwA) or they can be a separate part of the presmoothing models (labelled PSwoA). If the anchor scores are incorporated directly into the propensity-score model, the resulting presmoothing log-linear models, which we call "inner models", are obtained:

$$\log(p_{jl}) = \beta_0 + \sum_{i=1}^{T_r}\beta_{x,i}(x_j)^i + \sum_{k=1}^{T_e}\beta_{e,k}(e_l)^k + \sum_{d}^{T_{xe}}\sum_{d'}^{T_{ex}} + \beta_{xe,dd'}(x_j)^d(e_l)^{d'}$$

If the anchor scores instead are incorporated separately in the log-linear models, we obtain outer models:

$$\log\ P\left(X = x_j, e(\boldsymbol{D}_{Xl'}) = e(\boldsymbol{d}_{Xl'}), A = a_l\right) = \log(p_{jll'}) = \beta_0 + \sum_{i=1}^{T_r}\beta_{x,i}(x_j)^i +$$
$$\sum_{k=1}^{T_a}\beta_{a,k}(a_l)^k + \sum_{k=1}^{T_e}\beta_{e,k}(e_{l'})^k + \sum_{d}^{T_{xe}}\sum_{d'}^{T_{ex}}\beta_{xe,dd'}(x_j)^d\,(e_{l'})^{d'} + \sum_{c}^{T_{ea}}\sum_{c'}^{T_{ae}}\beta_{ea,cc'}(e_{l'})^c\,a_l^{c'} +$$
$$\sum_{i=1}^{T_{xa}}\sum_{k=1}^{T_{ax}}\beta_{xa,dd'}\,x_j^d a_l^{d'}.$$

The obtained models are then used to estimate the score probabilities when performing kernel equating.

An alternative approach to presmoothing is the EM-based log-linear method proposed in Liou (1998), which integrates test scores, anchor items, and group membership into a unified model. This approach explicitly accounts for ignorable and nonignorable missing-data mechanisms. While our method shares a conceptual foundation with this framework, it differs in that we use the propensity score $e(\boldsymbol{D})$ as a scalar balancing measure rather than modelling group membership effects directly. This allows for a flexible incorporation of background covariates while maintaining the benefits of log-linear smoothing

Empirical study

Data from the college admissions test SweSAT was used to illustrate the proposed extension of using propensity scores together with information from anchor tests within the kernel equating framework. SweSAT contains 160 multiple-choice, binary scored items, comprising a verbal section and a quantitative section of 80 items each. The two sections are equated separately. The test takers were also administered either an external 40 items (verbal or quantitative) anchor

test or 40 (verbal or quantitative) try-out items. Typically, the SweSAT is given twice a year to between 28,000 and 60,000 test takers, and about 2,000 test takers receive the 40 items quantitative anchor test. Before using anchor tests when equating the scores, the equating was done by using a set of covariates as described in Lyrén and Hambleton (2011). Although anchor tests are available nowadays, covariates are still of interest when equating the SweSAT as the empirical covariate distributions are not necessary the same at different administrations. Note that currently, when equating the SweSAT, several equating methods are used in practice including the KCE, KPSE and PS equating methods.

We used four administrations of the quantitative section as well as the quantitative anchor test to present two scenarios. For each scenario, we used the same covariates that are recorded and used in past administrations (Altintas & Wallin, 2021, Bränberg, et al 1990, Wallin & Wiberg, 2019, 2023) and the fact that we had access to them. Descriptive statistics of the verbal test scores (range 0–80), age, highest attained education and sex are given in Table 1 and 2. The verbal SweSAT test scores were grouped into four strata based on previous studies and analyses: [0–32], [33–43], [44–55], and [56–80]. Age was grouped into four strata: [0–20], [21–24], [25–29], [30–oldest], which is like Wallin and Wiberg (2019, 2023) except that we merged the two age categories with few test takers into a single highest age range (30-39 and 40-oldest). Highest attained education (Educ) was grouped into six strata which is reasonable from the Swedish school system: [9y; 9 school years], [AE; Adult education], [G2; 2 years upper secondary school], [G34: 3-4years upper secondary school, [2yC; 2 years of college], [m2yC; more than 2 years of college]. The quantitative anchor test was also used.

We assumed that we always equated a new test form X to an old test form Y. In scenario 1, we equated two test forms which had very different empirical distributions with respect to sex, age and education compared with all other administrations (see first two rows of Table 1). The reason was that test X1 was administered during a covid year, and it was equated to a test form given before Covid-19. The SweSAT is highly affected by the Swedish unemployment rate, as more test takers want to apply for university if they lose their jobs. The unemployment was higher during covid than the years before the pandemic. In the second scenario, we equated two administrations which had similar empirical distributions with respect to sex, age, and education (row 3 and 4 in Table 1) and the test forms were not administered during the pandemic. KPSE was used to equate the test forms when propensity scores were used.

In each scenario we compared the following method and designs: 1) NEC design with anchor test within the propensity score model (PSwA), 2) NEC design with propensity scores but

anchor outside the propensity score model (PSwoA), 3) propensity scores with a NEC design without anchor information (PS), 4) NEAT design with KCE, 5) NEAT design with KPSE.

Propensity scores were obtained with logistic regression using all covariates including the anchor test in 1), and all covariates excluding the anchor test in 2) and 3). The estimated propensity scores from the fitted model were divided into several strata according to the percentiles. The propensity score models were assessed by checking the covariate balance in the strata using the absolute standardized mean difference (ASMD) in which a difference of less than 0.1 indicate good balance (Austin, 2008). The AMSD is defined as

$$\text{ASMD} = \left| \frac{\mu_D^{(T)} - \mu_D^{(C)}}{\sqrt{\frac{[\sigma_D^{2(T)} + \sigma_D^{2(C)}]}{2}}} \right|,$$

where $\mu_D^{(T)}$ and $\mu_D^{(C)}$ are the means of test form X (treatment) and test form Y (control) for covariate D and $\sigma_D^{2(T)}$ and $\sigma_D^{2(C)}$ are their respective variances. We chose to use the number of strata so that this was achieved for as large fraction of strata as possible for every covariate. In our study, this was achieved with 13 strata. The average ASMD for the used covariates when anchor scores were within the propensity scores ranged from 0.02 (Gender) to 0.20 (Anchor) and when the anchor test scores were outside the propensity scores the range was 0.02 (Educ) to 0.15 (Age).

The Bayesian information criterion (BIC, Schwarz, 1978) was used to choose parametrization of the log-linear models in the presmoothing step as it has been shown to have a high selection accuracy for bivariate smoothing (Moses & Holland, 2010). The following log linear models were chosen for KCE and KPSE: X^3, A, AX, AX^2. For PS without anchor (PS) and PS with anchor inside (PSwA): X^3, ps^2, psX and for PS with anchor outside (PSwoA): X^3, ps^2, A, psX, psA. Note that, X^3 means that all lower terms are also included in the model, i.e. in this case also X^2 and X. The Gaussian kernel was used in the continuization step as that is used when kernel equating methods are used to equate the SweSAT.

Table 1

Descriptive statistics of the four administrations in total and for the anchor test groups

	Sex		Age					Educ						Verbal test scores			
Total																	
Adm	N	F	-20	21-24	25-29	30-39	40-	9y	AE	G2	G4	2C	m2C	0-33	34-44	45-55	56-80
Y1	55,072	52	58	23	10	6	2	3	1	3	76	9	5	15542	14290	12605	12635
X1	28,165	56	45	30	13	9	3	2	2	5	70	12	6	6877	7047	7120	7121
Y2	39,246	53	62	21	9	6	3	2	2	3	79	8	5	8501	10011	9312	10522
X2	58,990	52	63	23	7	5	2	2	1	3	81	7	3	15664	15403	13736	12572
Anchor																	
AY1	1578	53	52	24	13	8	2	3	2	4	74	10	5	500	431	355	292
AX1	1299	55	44	28	15	11	3	2	2	6	75	9	4	284	349	337	329
AY2	1727	53	63	21	9	5	2	1	1	4	81	8	5	335	476	448	468
AX2	2615	50	63	25	6	4	1	1	1	1	82	8	3	667	697	711	540

Adm = Administration, N = Number of test takers, F = Female percentage, 9y=9 school years, AE= Adult Education, G2=2 years of upper secondary school, G4 = 3-4 years of upper secondary school, 2C = 2 years of college, m2C = more than 2 years of college. AY1, AX1, AY2, AX2 = The anchor test forms given at the same administration as test forms Y1, X1, Y2 and X2.

Summary statistics including correlation are given in Table 2. Note that some of the covariates are quite similar over the four administrations, however for education it differs substantially. The means differed considerably, and the standard deviations differed a lot in scenario 1.

Table 2

Mean, standard deviation (SD) and correlation of the four administrations used in the two scenarios in the empirical study

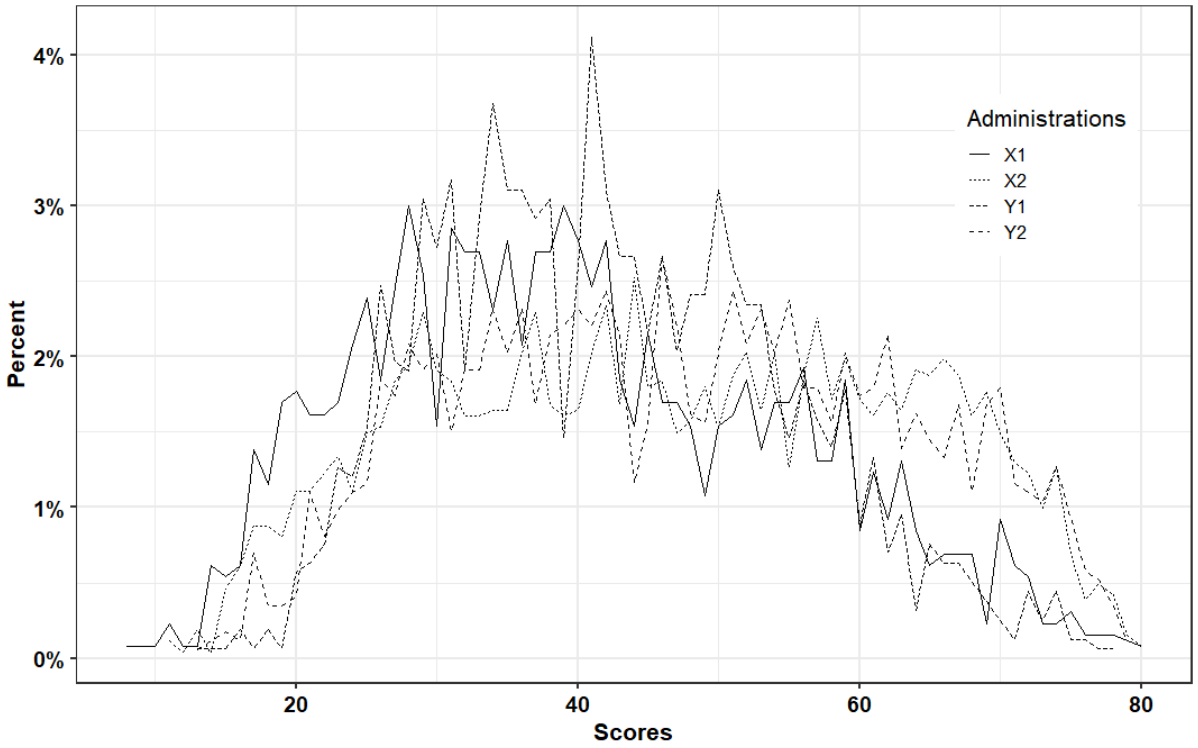
	Mean	SD	Correlation				
Scenario1			Sex	Age	Educ	Verb	A
Y1	43.79	12.54	-0.23	-0.16	0.13	0.57	0.84
X1	41.17	15.48	-0.25	-0.16	0.19	0.55	0.89
AY1	19.55	7.59	-0.24	-0.14	0.17	0.54	-
AX1	21.06	7.62	-0.26	-0.10	0.16	0.53	-
			Correlation				
Scenario 2	Mean	SD	Sex	Age	Educ	Verb	A
Y2	46.27	15.23	-0.22	-0.15	0.15	0.58	0.90
X2	45.05	16.25	-0.23	-0.17	0.11	0.55	0.92
AY2	21.75	7.93	-0.25	-0.10	0.16	0.56	-
AX2	22.75	8.14	-0.24	-0.12	0.12	0.56	-

SD= Standard deviation, Educ = maximum education, Verb = Verbal test scores, A = anchor test scores. AY1, AX1, AY2, AX2 = The anchor test forms given at the same administration as test forms Y1, X1, Y2 and X2. The correlation for the variable Sex is point-biserial correlation and Spearman correlation for Educ and Age.

Figure 1 displays the four test score distributions, and it is clear from both the mean and SD in Table 2 and Figure 1 that the test distributions are quite different, especially in the mid score range.

Figure 1

Test score distributions for both scenario 1 and scenario 2



To evaluate the equating methods used in the empirical study we used the same measures as Wallin and Wiberg (2019, 2023) used in their empirical studies, i.e., difference between the equated score and the raw score, and the SEE. The empirical study was carried out in R with the package *kequate* (Andersson, Bränberg, & Wiberg, 2013). To use propensity scores using *kequate*, one can simply replace the function call for the anchor with a call to the estimated and stratified propensity scores.

Results from the empirical study

The first row in Figure 2 illustrates the difference between equated scores and raw scores and the second row illustrates the equating transformations for the two scenarios when either NEAT design is used (KPSE and KCE) or NEC design with propensity scores is used (PS), or NEC design with anchor test within propensity scores (PSwA) or NEC design with propensity scores but anchor outside (PSwoA). The differences between equated scores and raw scores are much larger for lower test scores and are especially large in scenario 1. Clearly the equating transformations are quite similar, especially in scenario 2 regardless of the method used. In scenario 1, PS and PSwoA differed most from the other equating transformations.

Figure 2

Difference between equated scores and raw scores (first row) and the equating transformations (second row) for scenario 1 (left) and scenario 2 (right)

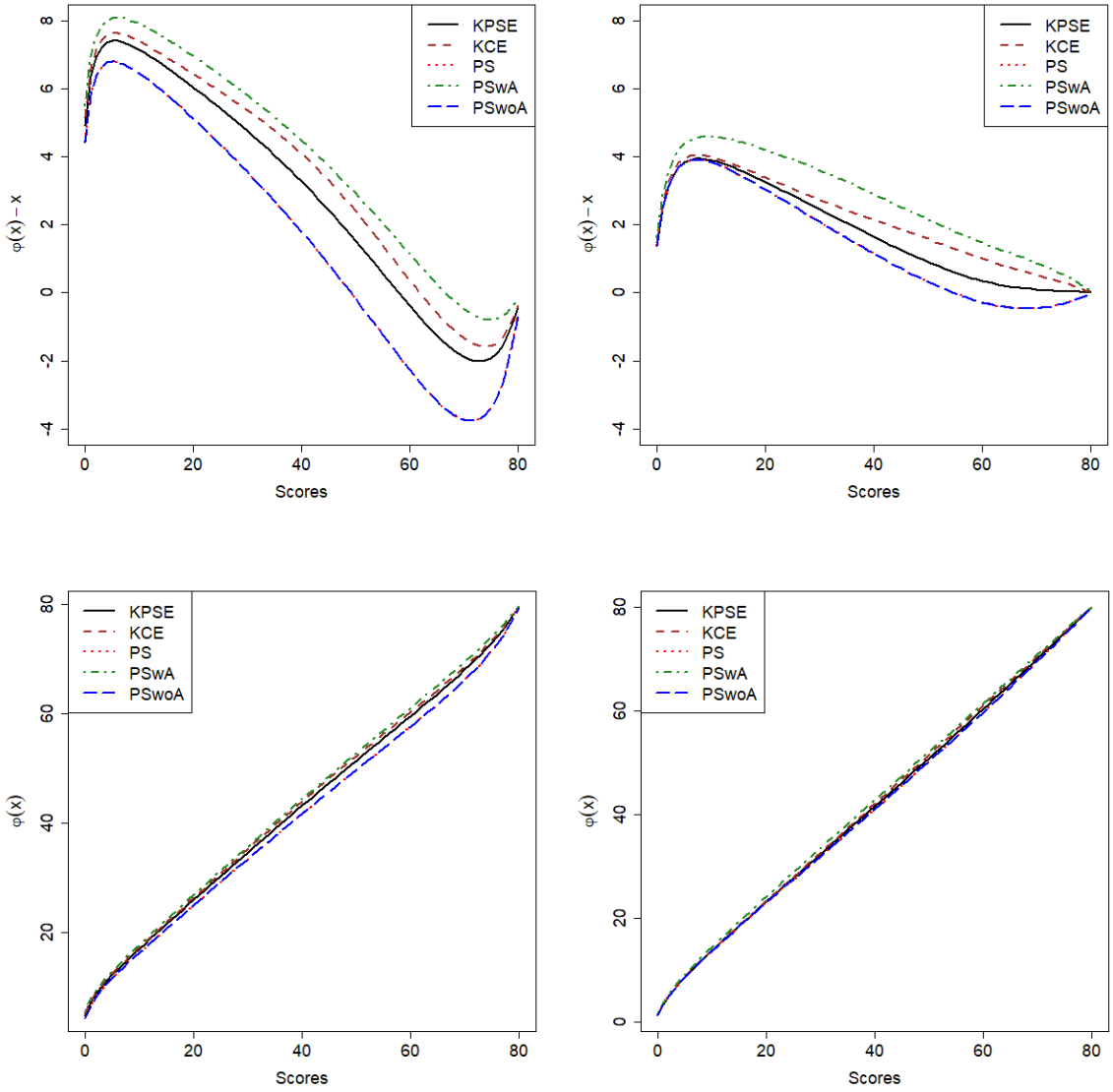


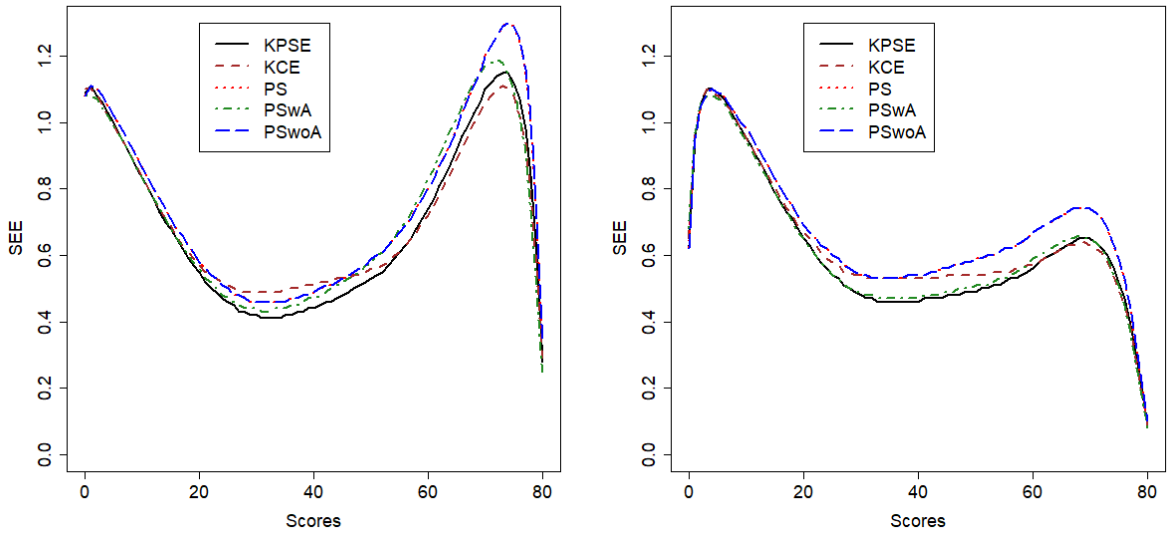
Figure 3 displays the SEE, and from this figure it is evident that when anchor test scores are included in the propensity scores the SEE is lower than if the anchor test scores are modelled as a separate term in the loglinear presmoothing models. The SEE is much higher in the low and high score range for both scenarios but as expected much lower in scenario 2. It is also interesting to note that SEE for KCE is higher in the mid score range than for the methods using covariate information in both scenarios. To demonstrate how loglinear presmoothing works, we added histograms comparing the distributions of non-smoothed and smoothed Form X scores,

as well as comparisons between the methods, and they can be seen in the Appendix A, figures A1 and A2.

Summing up, when there are significant differences in the test distributions (Scenario 1), the SEE and the discrepancies between equated scores and raw scores were larger than when the score distributions were more similar (Scenario 2). Also, when anchor test scores are incorporated within the propensity score estimation, we obtained lower SEE compared to when they are treated as separate covariates.

Figure 3

SEE for scenario 1 to the left and scenario 2 to the right



Simulation study

To be able to examine several different conditions we conducted a simulation study in which we varied number of test takers, anchor items, and the correlation level between the covariates and the test scores. In addition, the abilities of the test taker groups, and the difficulty of the anchor test were varied. In the following, the simulation design and the evaluation measures re described. For each simulation scenario, 500 replications were used. First, we summarize the scenarios considered, before describing how the simulated data was generated.

- Two populations, P and Q, were generated, each with a population size of 200,000 test takers.
- A subset of either 1,000 or 2,000 test takers was sampled for each replication. A regular test length of 80 and a varying anchor test length of either 20 or 40 were used.
- Low and moderate correlations between the covariates and the test scores were considered.

With two sample sizes, two anchor test lengths, and two correlation scenarios, we had 32 scenarios in total (see Table 3). Next, a description on how the data was generated is given.

Table 3 *Scenarios (S) in the simulation study.*

S	P=Q	P	A_b	A_{max}	weak corr	moderate corr	N
S1	X			20	X		1000
S2	X			20		X	1000
S3	X			40	X		1000
S4	X			40		X	1000
S5		+		20	X		1000
S6		+		20		X	1000
S7		+		40	X		1000
S8		+		40		X	1000
S9	X		+	20	X		1000
S10	X		+	20		X	1000
S11	X		+	40	X		1000
S12	X		+	40		X	1000
S13		+	+	20	X		1000
S14		+	+	20		X	1000
S15		+	+	40	X		1000
S16		+	+	40		X	1000
S17	X			20	X		2000
S18	X			20		X	2000
S19	X			40	X		2000
S20	X			40		X	2000
S21		+		20	X		2000
S22		+		20		X	2000
S23		+		40	X		2000
S24		+		40		X	2000
S25	X		+	20	X		2000
S26	X		+	20		X	2000
S27	X		+	40	X		2000
S28	X		+	40		X	2000
S29		+	+	20	X		2000
S30		+	+	20		X	2000
S31		+	+	40	X		2000
S32		+	+	40		X	2000

Note. P=Q: P and Q have similar ability, P = Group P is more (+) capable, A_b = Anchor test form is more (+) difficult than the regular test forms, A_{max} = number of anchor items, N = sample size.

Note that the simulation condition S23 (or S7 with a smaller sample) is the closest to the empirical study. From the anchor test results presented in Table 2, it is evident that X samples

performed better than Y samples. Correlations between the scores and covariates were weak. In operational settings, anchor sample sizes ranged from 1000 to 2000, with the majority being closer to the upper end of that range.

Data-Generating Process

Two matrices, $\boldsymbol{P}$ and $\boldsymbol{Q}$, were initialized with dimensions corresponding to the population size (N) and the total number of items plus covariates ($M + L_1 + L_2 + L_3$), where M is the total number of items (regular test items and anchor items) and $L_1 = 3$, $L_2 = 4$, and $L_3 = 5$ represent the number of categories for each covariate, respectively. While we used categorized covariates in this study to match our empirical data conditions, the propensity score equating method is flexible and can accommodate continuous covariates as well. Researchers with access to continuous variables such as age or test scores may choose to use them directly in the propensity score estimation without categorization. The choice between categorical and continuous covariates should be guided by data availability and the specific research context.

For each test taker in the population, item responses were generated using the item response theory (e.g. van der Linden, 2018) logistic function:

$$P(Y_{ij} = 1|\theta_i) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}},$$

where θ_i represents the latent ability of test taker i, and a_j and b_j represents discrimination and difficulty for item j, respectively. Latent abilities for populations P and Q were drawn from normal distributions $N(0,1)$ and $N(0.2,1)$, respectively. The item difficulty parameters were drawn from a $N(0,1)$ distribution, and the item discrimination parameters were drawn from a $U(0.5,2)$ distribution. All item parameters were drawn independently of each other. The binary responses were then determined by comparing the logistic probability to a uniform random variable:

$$Y_{ij} = \begin{cases} 1, & \text{if } \dfrac{1}{1 + \exp\left(-a_j\left(\theta_i - b_j\right)\right)} > U \\ 0, & \text{otherwise} \end{cases}$$

where $U \sim U(0,1)$. The sum scores for the regular test items and the anchor items from the generated item responses were computed for each test taker. The covariates for each test taker were generated similarly. For each test taker i and covariate k,

$$P(C_{k,i} = 1|\theta_i) = \frac{1}{1 + e^{-a_{C_k}\left(\theta_i - b_{C_k}\right)}},$$

where $b_{C_k} \sim N(0, 1)$ and the covariate item parameters a_{C_k} were generated according to different correlation structures:

- Low correlation setting: $a_{C_k} \sim U(0.1, 0.5)$
- Moderate correlation setting: $a_{C_k} \sim U(0.5, 1.5)$

Lastly, we calculated the sum score of each covariate, thus creating three categorical covariates.

Population model and four estimators

We examined a population model and four alternative estimators. For each scenario and estimator, the best-fitting log-linear models were selected separately using the Akaike information criterion (AIC; Akaike, 1974), BIC, and the likelihood ratio test (LRT; Haberman, 1974a, 1974b). This resulted in at most six unique models – one for (X, A) and one for (Y, A) per criterion. In the second step, all possible combinations of these model pairs were evaluated, and the pair that minimized the average SEE across test scores was chosen, following the approach suggested by Wallin and Wiberg (2024).

Population model: Using the population-level data, propensity scores were estimated using a logistic regression model with the covariates as predictors. Test takers were stratified into 15 groups based on these propensity scores. Equating was thereafter performed using KPSE and KCE methods.

Common procedure for Estimators 1-3: For all propensity score estimators, test takers were stratified into 15 groups based on propensity scores, and the strata acted as predictors in the log-linear model together with the score variables. The estimators differ in their propensity score specification:

Estimator 1: Equating with propensity score that includes only covariates (PS)

- Propensity scores were estimated using covariates only.
- Equating was performed using KPSE and KCE methods.

Estimator 2: Equating with propensity scores that includes both covariates and anchor scores (PSwA)

- Propensity scores were estimated including both covariates and anchor items.
- Equating was performed using KPSE and KCE methods.

Estimator 3: Equating with anchor score outside of propensity score (PSwoA)

- Propensity scores were estimated only with covariates.
- A three-dimensional contingency table was created for the sum score, propensity score strata, and anchor score.
- Equating was performed using the PSE method.

Estimator 4: NEAT Equating (KCE/KPSE)

- Kernel equating with the NEAT design using both KCE and KPSE was conducted as a baseline comparison.

Evaluation measures

To evaluate the equating transformations, we used four evaluation measures. We examined bias, over R replications

$$\text{Bias}(\hat{\varphi}_Y(x_i)) = \frac{1}{R}\sum_{g=1}^{R}(\hat{\varphi}_Y^{(g)}(x_i) - \varphi(x_i)),$$

where $\varphi(x_i)$ is the true equating transformation. The true equating transformations for each estimator (population-level KPSE and KCE) were defined based on the true propensity scores, which were calculated using a logistic function of the anchor scores and the covariates. The NEAT KCE estimators were compared against the identity function, a valid procedure due to the data-generating process with difficulty parameters drawn from the same distribution (Laukaityte & Wiberg, 2024, Leoncio, Wiberg & Battauz, 2023). The equating transformations were then derived from log-linear models fit to the population-level frequency tables of test scores and categorized propensity scores. This setup ensured that the equating transformations reflected the true relationship between the test scores, the covariates and the anchor scores.

We examined the SEE from Equation 4, and the root mean squared error (RMSE),

$$\text{RMSE}(\hat{\varphi}_Y(x_i)) = \sqrt{\frac{1}{R}\sum_{g=1}^{R}(\hat{\varphi}_Y^{(g)}(x_i) - \varphi(x_i))^2},$$

and the standard error (SE)

$$\text{SE}(\hat{\varphi}_Y(x_i)) = \sqrt{\frac{1}{R-1}\sum_{g=1}^{R}(\hat{\varphi}_Y^{(g)}(x_i) - \bar{\varphi}_Y^{g})^2},$$

where and $\bar{\varphi}_Y^{g} = \frac{1}{R}\sum_{g=1}^{R}\varphi_Y^{(g)}(x_i)$. The simulation study was carried out in R with the R package *kequate* (Andersson, Bränberg & Wiberg, 2013). The used code can be found on the following github: https://github.com/gabrieltwallin/Equating_anchor_PS.

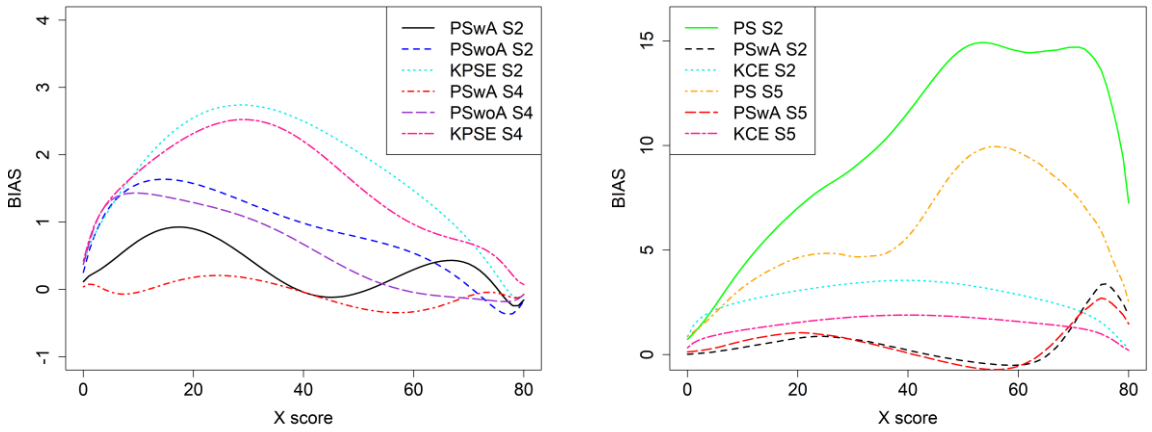
Results from the Simulation Study

In the simulation study, in addition to varying sample sizes, anchor test lengths, and correlation strength, we also varied the abilities of the test-taker groups and the difficulty of the anchor test. Note, in all figures in this section the left figures are based on the KPSE estimator, and the right figures are based on the KCE estimator. Figures 4 and 5 present the results for the baseline case where the groups had similar abilities, and the difficulty of the regular test forms and the anchor test form were comparable. The difference between the two figures is the strength of the correlation between the covariates. When the correlation between the covariates was moderately strong (see Figure 4a, b), the differences in bias between the studied equating methods were larger compared to when the correlation was weak (see Figure 8a, b), especially for KPSE. However, the differences in bias between the various anchor test lengths were more pronounced when the correlation was weak. For KCE, the differences in bias across different correlation strengths or anchor test lengths were small.

The main differences in SEE were observed between the different methods, with the smallest SEE occurring when the anchor score was outside of the propensity score (PSwoA) and the anchor test consisted of 40 items for KPSE (see Figure 4c and 5c), and for the NEAT design when using KCE (see Figure 4d and 5d).

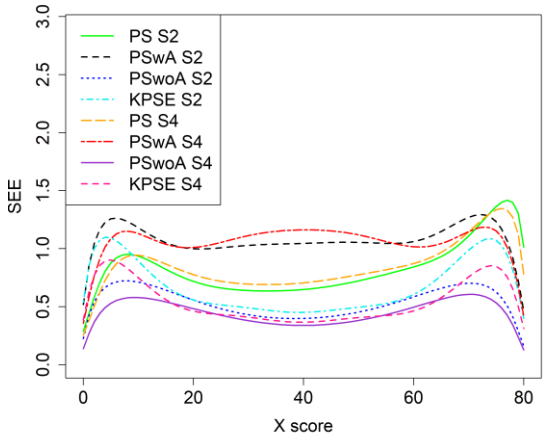
Figure 4

Bias (a and b) and SEE (c and d) for the baseline case when correlation between covariates was moderate strong and the size of an anchor test form was either 20 items (S2) or 40 items(S4)

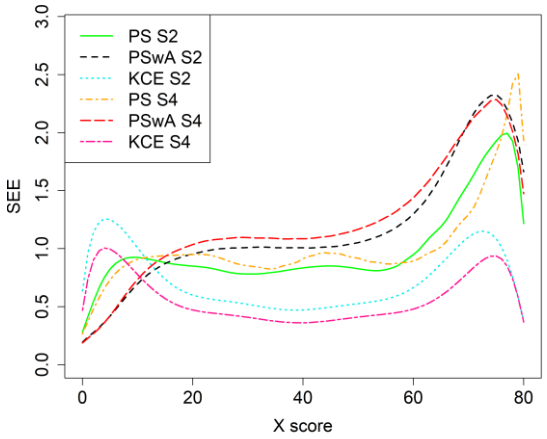


a)

b)



c)



d)

Note that for KPSE, we do not show the bias results for the PS method, as the bias is so large (see Appendix Figures B1a and B2a) that it obscures the differences between the other methods. For KCE, we used an identity function as a criteria function when evaluating bias for NEAT KCE. Otherwise, it resulted in large bias values similar to the PS results for KPSE (see Appendix Figures B1b and B2b).

Figure 5

Bias (a and b) and SEE (c and d) for the baseline case when correlation between covariates was weak and the size of an anchor test form was either 20 items (S1) or 40 items(S3)

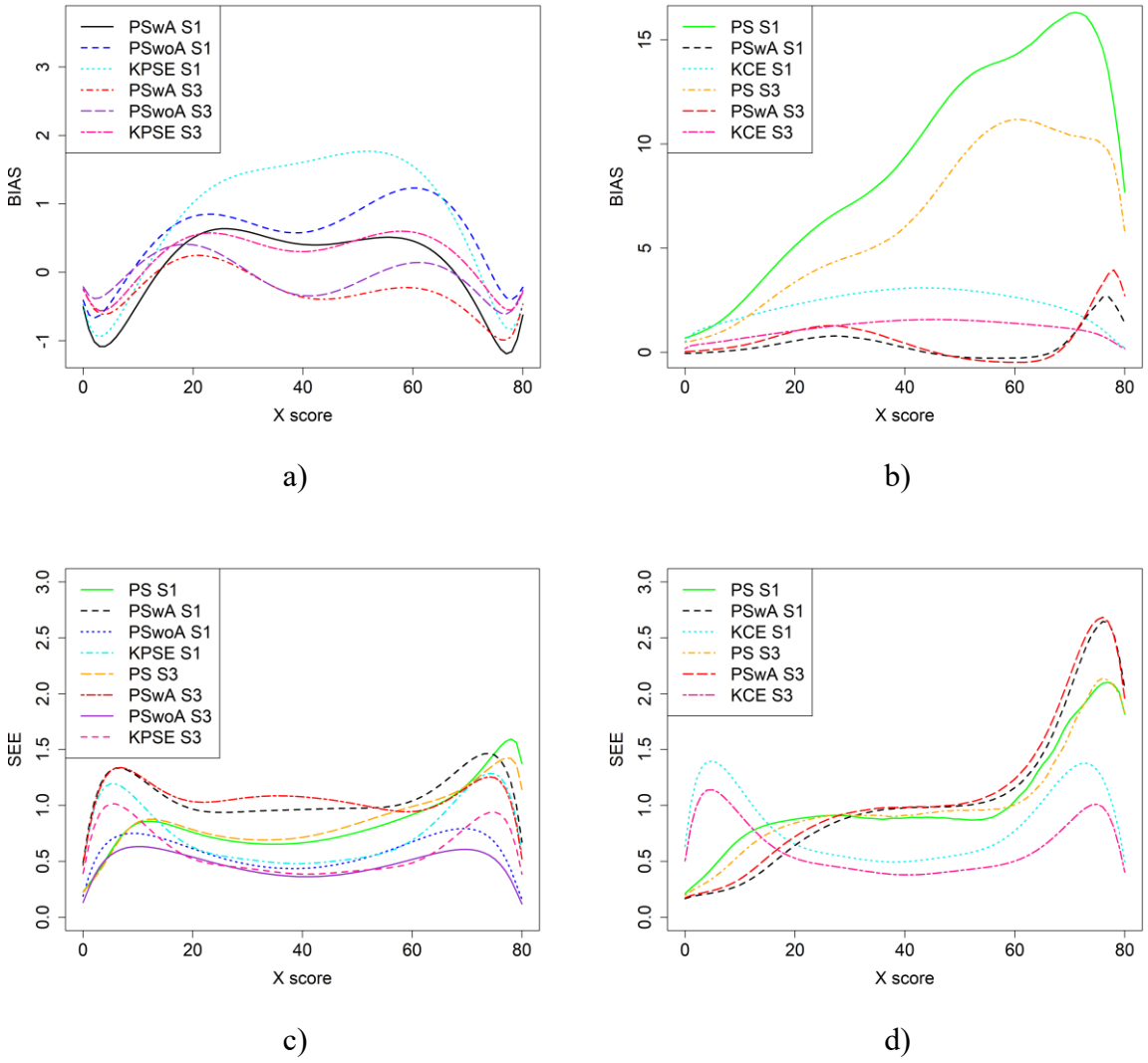


Figure 6 displays the RMSE and SE for the baseline case in Figure 4. As their results are similar to the bias and SEE figures – we draw the same conclusions from them. For subsequent scenarios, we have therefore omitted RMSE and SE figures, but these can be obtained upon request from the corresponding author.

Figure 6

RMSE (a and b) and SE (c and d) for the baseline case when correlation between covariates was moderate and the size of an anchor test form was either 20 items (S2) or 40 items(S4)

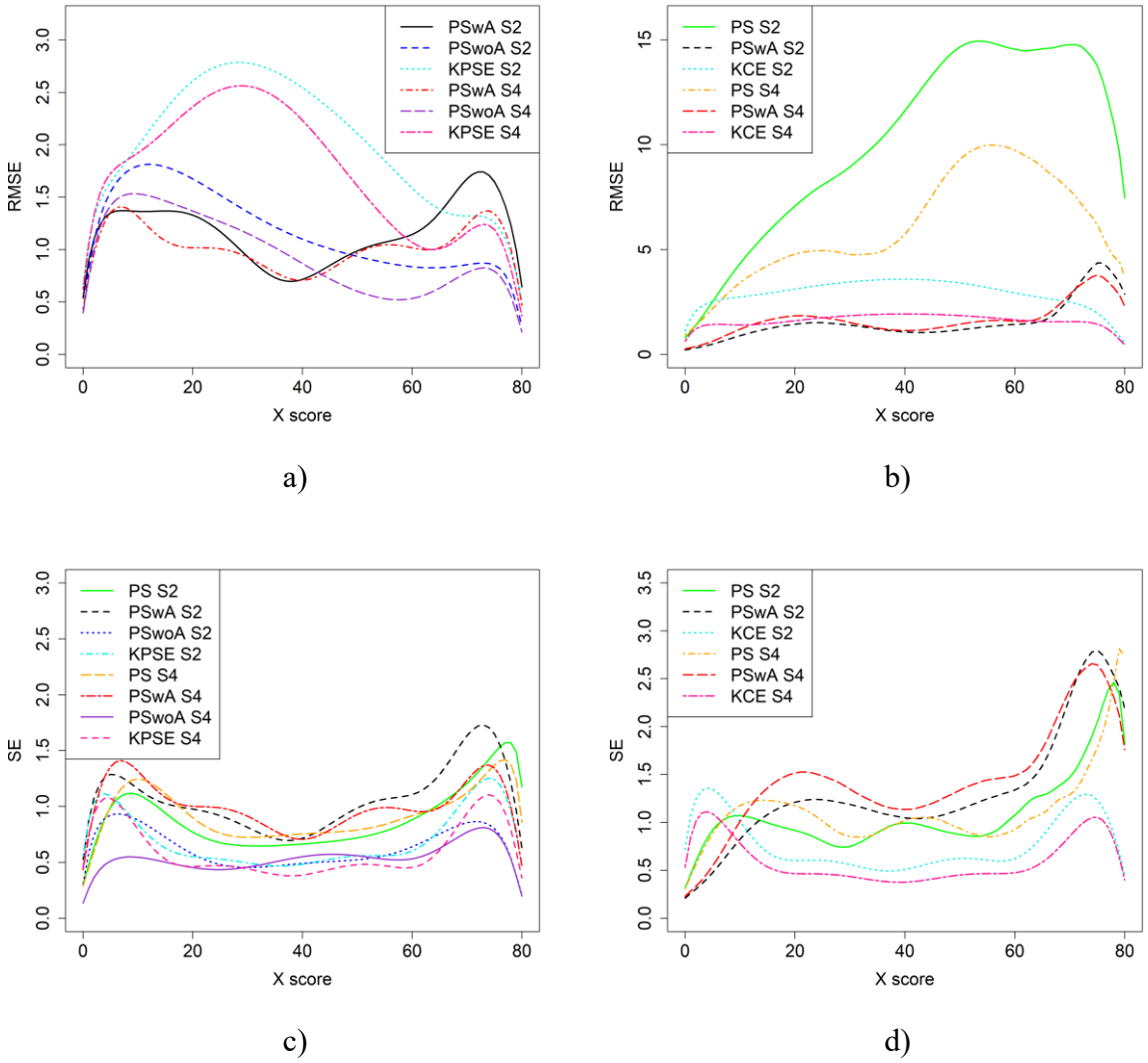


Figure 7 presents the results for bias and SEE for the baseline case for different sample sizes: 1000 (S2) and 2000 (S18). The bias only indicated minimal or no differences. The sample size, however, impacted the SEE results, with the largest differences occurring for equating with propensity scores (PS) and with propensity scores that included both covariates and anchor scores (PSwA).

Figure 7

Bias (a and b) and SEE (c and d) for the baseline case when correlation between covariates was moderate and the sample size N was either 1000 (S2) or 2000 (S18)

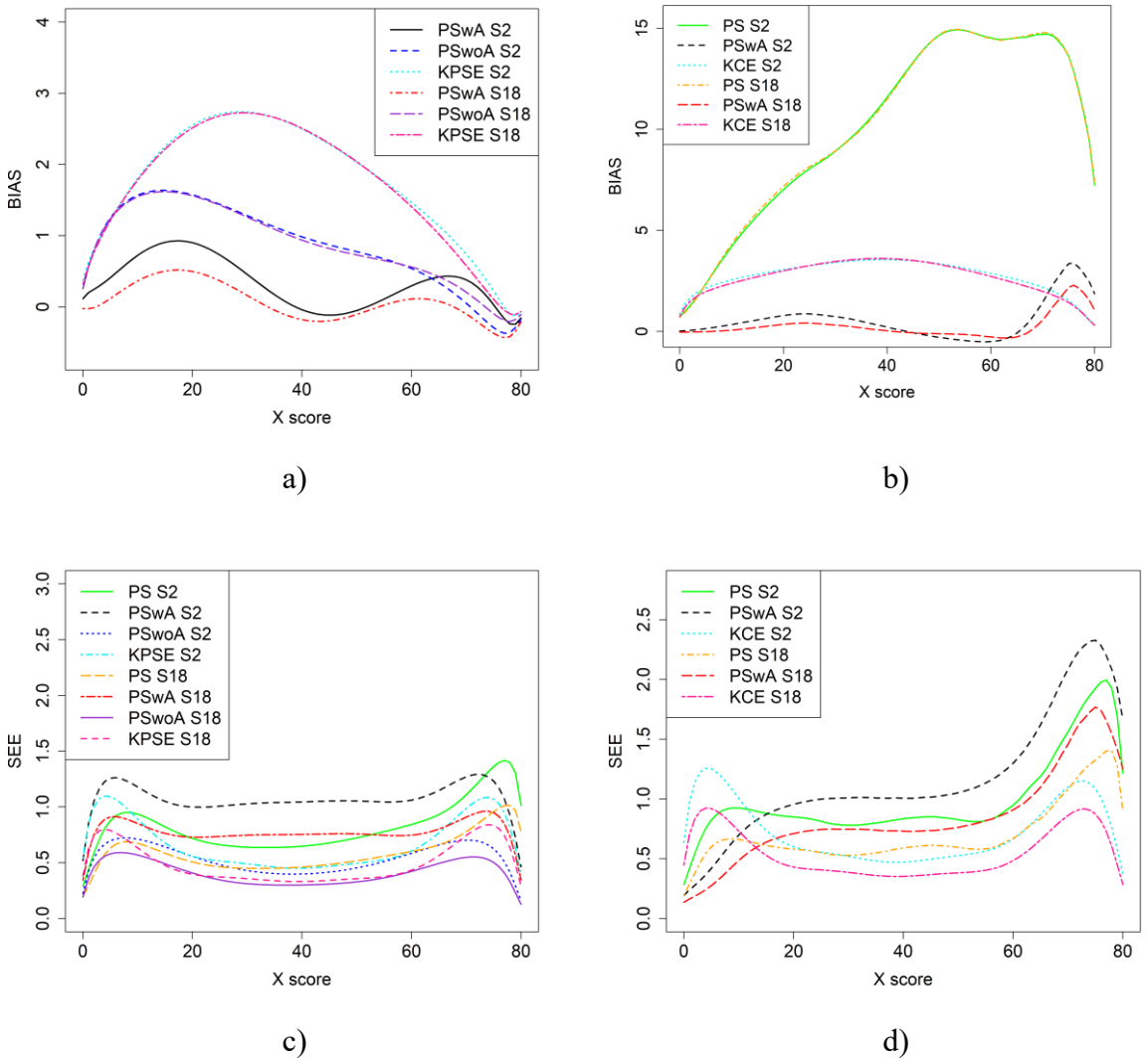
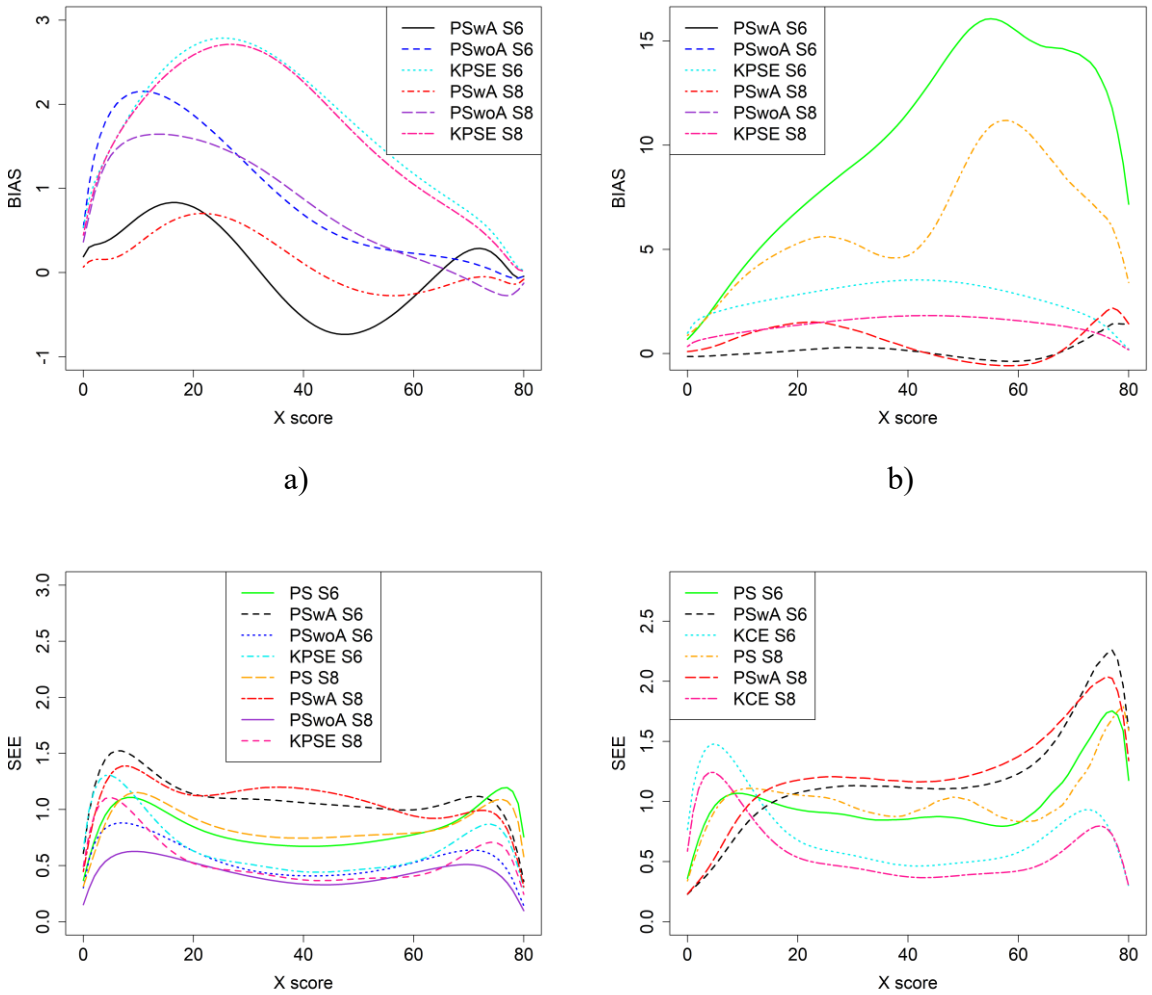


Figure 8 displays result similar to those shown in Figure 4. However, in this case, the scenarios involve one group with average ability and another with higher ability. SEE values for KPSE are nearly identical to those in the baseline case shown in Figure 4. For KCE, when the correlation between the covariates is moderate, SEE values are slightly lower at the high scores, especially for PswA and PS with longer anchor test (see Figure 10b) compared to the baseline case. In contrast, when correlation is weak, SEE values are higher at the high scores for these same methods (see Appendix Figure B3). The differences in group abilities had a slightly greater impact on bias values, particularly for KPSE compared to KCE. When the correlation

between the covariates was weak, bias increased for the lower scores for KPSE equating, unlike in the baseline case.

Figure 8

Bias (a and b) and SEE (c and d) for groups differing in ability when correlation between covariates was moderate and the size of an anchor test form was either 20 items (S6) or 40 items(S8)



If the anchor test form is more difficult than the regular test forms, the bias results change significantly, especially when the correlation between the covariates is moderate (see Figure 9). The largest changes in bias are observed for equating methods using propensity scores. Interestingly, when the correlation between the covariates was weak, the bias results for KCE

were similar to the baseline case (see Figure 4). The difficulty of the anchor test had only a minor effect on SEE values.

Figure 9

Bias (a and b) and SEE (c and d) for groups similar in ability when anchor test form was more difficult than the regular test forms, correlation between covariates was moderate and the size of an anchor test form was either 20 items (S10) or 40 items(S12).

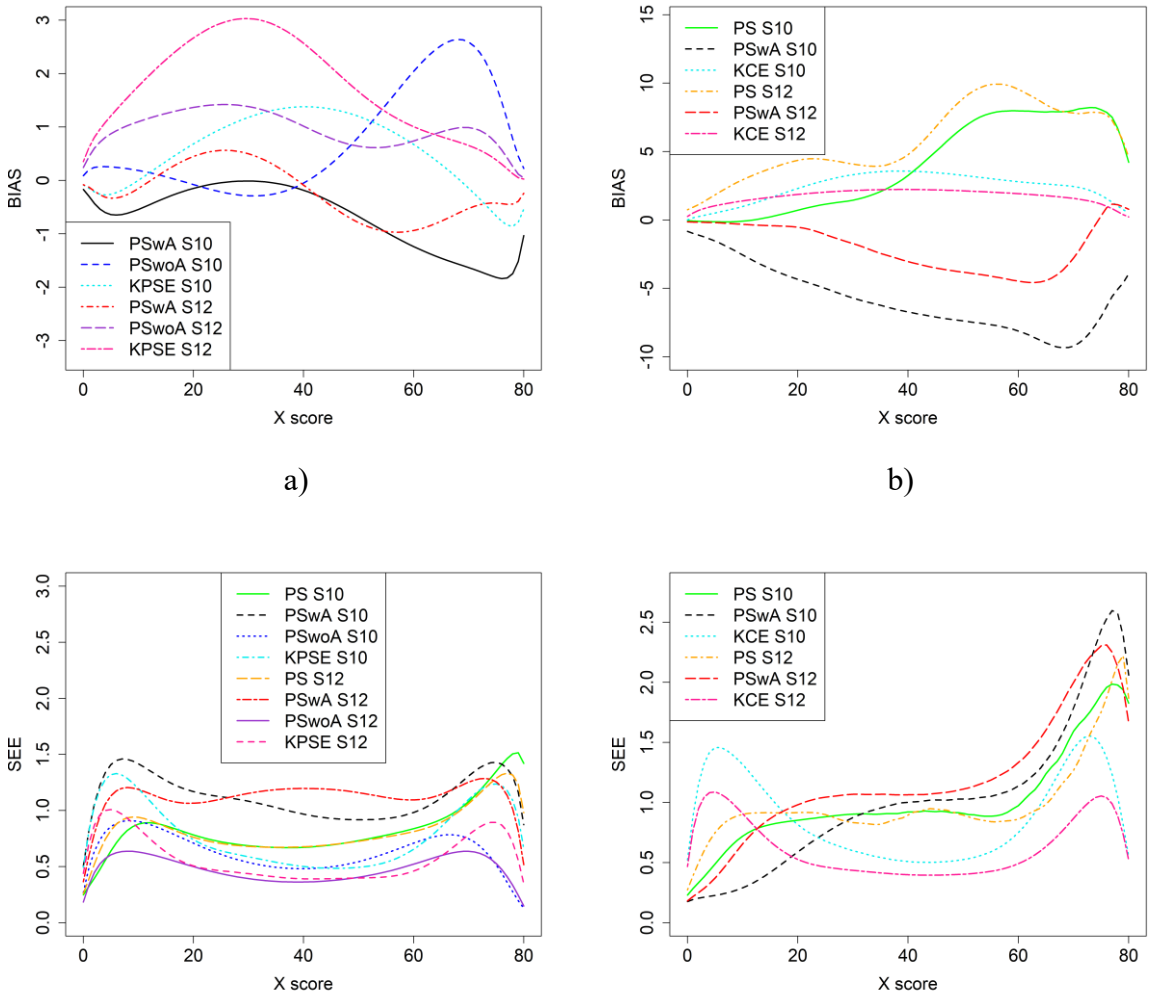
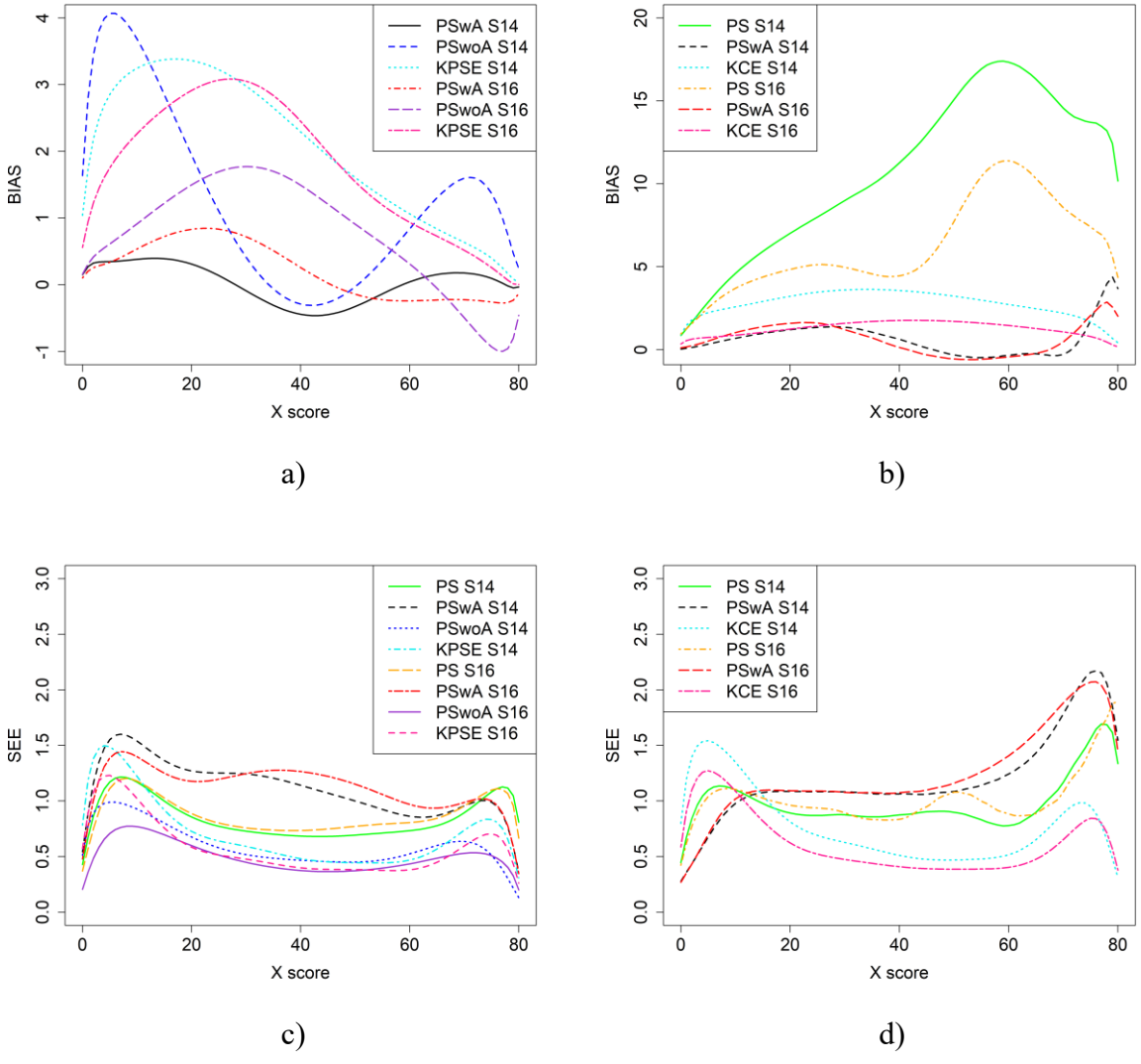


Figure 10 displays the equating results when the anchor test form is more difficult than the regular test forms and one group has higher ability than the other. For the KPSE based methods, PSwoA had the lowest SEE and PSwA had in general the lowest bias. The greatest impact on bias values was seen for the KPSE methods, compared with the KCE methods.

Figure 10

Bias (a and b) and SEE (c and d) for groups differing in ability when anchor test form was more difficult than the regular test forms, correlation between covariates was moderate and the size of an anchor test form was either 20 items (S14) or 40 items(S16).



Discussion

The primary objective of this study was to propose and evaluate a novel approach that could be incorporated into generalized kernel equating. The proposed approach integrates propensity scores with anchor test scores. This approach was compared against two established methods: the NEAT design alone and the NEC design using propensity scores without anchor test scores. The empirical study showed that when test distributions exhibited significant differences (Scenario 1), the SEE and the discrepancies between equated scores and raw scores were more

pronounced. This finding aligns with the results of Laukaityte and Wiberg (2024). Notably, the use of propensity scores led to a lower SEE, consistent with the conclusions of Wallin and Wiberg (2019), although their study did not examine the combined use of anchor test scores, and propensity scores derived from covariate information. As expected, incorporating anchor test scores within the propensity score estimation resulted in a lower SEE compared to treating them as separate covariates. This suggests that the integration of anchor test information into propensity scores may enhance the precision of equating. This is also in line with the results of Kim and Walker (2021; 2022) who concluded that using both sample weights via MDIA and a short anchor produced the most accurate equating results.

From the simulation study, we concluded that when the correlation between the covariates was moderately strong, the differences in bias and RMSE between the methods were larger compared to when the correlation was weak, especially when using KPSE. The difference was also more pronounced when a shorter anchor test was used in conjunction with weak correlation. This is not surprising, as a shorter anchor test and weaker correlation yield less overall information. For KCE, the differences in bias across different correlation strengths or anchor test lengths were small. For KPSE, the smallest SEE and SE occurred when the anchor score was outside of the propensity score (PSwoA) and the anchor test consisted of 40 items. For the NEAT design, the smallest SEE and SE occurred when using KCE. This is expected, as more information about the test takers should yield a smaller error, as seen, for example, in Bränberg and Wiberg (2011), who examined observed score linear equating with covariates.

Varying the sample sizes had little effect on bias but did impact the SEE results, with the largest differences occurring when equating with propensity scores (PS) and with propensity scores that included both covariates and anchor scores (PSwA). In general, SEE was lower when anchor test scores were used as a separate covariate (PSwoA) compared to when they were included within the propensity score (PSwA). This is probably because treating the anchor scores as a separate covariate provides more information about the test takers than when the anchor scores are combined with other covariates within the propensity score.

When the ability between the groups differed, the SEE values for KPSE were nearly identical to those in the baseline case. This is in line with Lu and Guo (2018) who concluded that if the ability group difference were large, to use NEAT is preferred in terms of RMSE and bias, instead of using only information in background variables through PEG. Also, our conclusion to use background information together with anchor test information is in line with their conclusion of using PEG procedures based on background variables together with the anchor test to improve the equating. When group ability differences were small (baseline case), the

SEE were low when either PSwoA or NEAT (KCE/KPSE) were used. Although we used a different approach and used bias and SEE to evaluate, our result is in line with Luo and Gao (2018), who concluded that using only NEAT design compared with using PEG without an anchor test gave comparable results, in terms of bias and RMSE.

When the correlation was moderate, SEE values were slightly lower at the higher scores, especially for PSwA and PS with a longer anchor test when KCE was used. This result is contrary to the findings of Ricker and von Davier (2007), who concluded that a shorter anchor yields a larger bias for KCE compared to KPSE; however, they did not examine the effect of correlation. Note that when the correlation was weak, the SEE values are higher at the higher scores for the same methods. The differences in group abilities had a slightly greater impact on the bias values, especially for KPSE compared to KCE. This is in line with, for example, Puhan (2010) and Power and Kolen (2014), who concluded that CE is less affected by group differences. When the correlation between the covariates was weak, bias increased for the lower scores for KPSE equating, unlike in the baseline case. Luo and Gao (2018) conclusion that if the anchor test is weak (i.e. few items and low correlation), is like the conclusion here, i.e. that we can then improve the equating with background information. When the anchor test form is more difficult than the regular test form, the bias results change significantly, particularly when the correlation between the covariates is moderate. Notably, the bias is especially large when using propensity scores without anchor test information. A possible explanation is that the propensity scores diverge too much from the anchor test scores, though this requires further investigation.

In summary, when the anchor test form is more difficult than the regular test forms and one group has higher ability than the other, the bias was more affected when using KPSE methods compared to KCE methods. These results are consistent with those of Laukaityte and Wiberg (2024), who studied how differences in group abilities impact kernel equating methods. If one has access to covariates, it is advisable to include them in the presmoothing model, as this can reduce the SEE. When multiple covariates at different levels are available, using propensity scores is an effective way to incorporate a large amount of information. In our study, it was also evident that, in terms of bias and SEE, it is better to include the anchor scores as a standalone covariate rather than incorporating them into the propensity score model.

This study has some limitations. First, we included only binary-scored items; in the future, polytomously scored items and mixed-format tests incorporating information from covariates should be examined. For example, Wallmark, Josefsson, and Wiberg (2023) examined kernel equating in mixed-format tests. A second limitation is that we examined only a few

presmoothing models. Wallin and Wiberg (2020; 2024) have shown that the presmoothing model has a significant impact on the equating transformation; therefore, several other models should be explored in future research. Another limitation concerns the choice of covariates and future studies should investigate other covariates and their usefulness when equating test scores. Note, there is a trade-off between test takers performance and precision of the test depending on the design of the test. On one hand, to include an anchor test prolongs the testing time and thus makes test takers more fatigue, on the other hand more information about the test takers is collected when including an anchor test and thus the precision of the equating can be increased.

While our primary focus was on horizontal equating scenarios where test-taker ability distributions differ but test content is similar, our approach may also be applicable to vertical equating. In vertical equating, test forms are tailored for different school grades, introducing additional complexities in modelling ability differences. As suggested in prior work (Liou, 1998), nonignorable missing-data models may be more appropriate in such contexts. Our method could potentially be adapted for vertical equating by extending the log-linear model to include additional covariates representing developmental differences across grades. Exploring this extension remains an interesting possibility for future research. Furthermore, continuous propensity scores can be used directly as conditioning variables in equating without requiring stratification, similar to how anchor scores function in traditional equating designs.

Another limitation is that the current standard error estimation approach does not explicitly account for the covariance between the empirical distributions F and G in the synthetic population. While we follow the framework of Wallin & Wiberg (2019) which provides estimates for the variances within each distribution, incorporating the covariance component would provide more accurate standard error estimates for the equated scores. Future methodological work should address how this covariance can be systematically incorporated into the standard error calculations for propensity score equating methods.

Finally, given that adjusting test score scales is a practical issue in many large-scale assessments, we included an empirical study to address this problem. Our results suggest that utilizing information from covariates, when they are available and informative, can be beneficial. However, we emphasize that an anchor test should also be used if available.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Albano, A. & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, and sample size. *Applied Psychological Measurement. 43*(8)*,* 597-610, https://doi.org/10.1177/0146621618824855

Altintas, Ö. & Wallin, G. (2021). Equality of admission tests using kernel equating under the non-equivalent groups with covariates design. *International Journal of Assessment Tools in Education, 8*(4), 729–743.

Andersson, B., Bränberg, K. & Wiberg, M. (2013). Performing the kernel method of test equating using the package kequate. *Journal of Statistical Software, 55*, 1-25.

Austin, P. C. (2008). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety, 17*(12), 1202–1217.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661-3679.

Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education and age on test scores on the Swedish scholastic aptitude test. *Scandinavian Journal of Educational Research, 34*(3), 189–203.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440.

Dorans, N. J., & Feigenbaum,M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

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

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.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioural Statistics 40*, 254-273.

Häggström, J. & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement, 51*(2), 201-211.

INVALSI (2013). Rilevazioni nazionali sugli apprendimenti 2012-13. Technical report, INVALSI Publishing. Retrieved May 25, 2023, from www.invalsi.it/snvpn2013/rapporti/Rapporto_SNV_PN_2013_DEF_11_07_2013.pdf/

Kim, S., & Walker, M. E. (2021). *Comparisons among approaches to link tests using random samples selected under suboptimal conditions*. (Research Report No. RR-21-14). Educational Testing Service. http://doi.org/10.1002/ets2.12328

Kim, S., & Walker, M. E. (2022). Adjusting for Ability Differences of Equating Samples When Randomization Is Suboptimal. *Educational Measurement: Issues and Practice*, *41*(3), 26-37.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3, 97–104. https://doi.org/10.1207/s15324818ame0301_7

Laukaityte, I. & Wiberg, M. (2024). Impacts of differences in group abilities and anchor test features on three non-IRT test equating methods. *Practical Assessment, Research, and Evaluation. 29*(5), 1-23. https://doi.org/10.7275/pare.2020

Leoncio, W., Wiberg, M., & Battauz, M. (2023). Evaluating equating transformations in IRT observed-score and kernel equating methods. *Applied psychological measurement*, *47*(2), 123-140.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95.

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica, 8,* 669-690.

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics, 40*, 227-253.

Lu, R., & Guo, H. (2018). *A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking* (Research Report No. RR-18-08). *ETS Research Report Series.* https://doi.org/10.1002/ets2.12196

Lu, R., & Kim, S. (2021). Effect of statistically matching equating samples for common-item equating. (Research Report No. RR-21-02) *ETS Research Report Series*, https://doi.org/10.1002/ets2.12313

Lyrén, P.-E., & Hambleton, R. K. (2011). Consequences of violated the equating assumptions under the equivalent group design. *International Journal of Testing, 36*, 308–323.

Moses, T., Deng, W., & Zhang, Y-L. (2010). *The use of two anchors in the nonequivalent groups with anchor test (NEAT) equating*. ETS research report RR-10-23.

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

Ozsoy, S.N. & Kilmen, S. (2023). Comparison of kernel equating methods under NEAT and NEC designs, *International Journal of Assessment Tools in Education, 10*(1), 56-75.

Paek, I., Liu, J., & Oh, H. J. (2006). *Investigation of propensity score matching on linear/nonlinear equating method for the P/N/NMSQT* (Report SR-2006-55). Princeton, NJ: ETS.

Powers, S. J. (2010). *Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions.* PhD thesis, University of Iowa. http://ir.uiowa.edu/etd/875

Powers, S. & Kolen, M. J. (2014). Evaluating equating accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement, 51*(1), 39-56. https://www.jstor.org/stable/24018322

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54-75. https://doi.org/10.1111/

Quenette, M. A., Nicewander, W. A., & Thomasson, G. L. (2006). Model-based versus empirical equating of test forms. *Applied Psychological Measurement, 30*(3), 167–182.

Ricker, K. L., & von Davier, A. A. (2007). The impact of anchor test length on equating results in a nonequivalent groups design. *ETS Research Report Series*, *2007*(2), i-19. https://doi.org/10.1002/j.2333-8504.2007.tb02086.x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516–524.

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

Stage, C., & Ögren, G. (2004). *The Swedish Scholastic Assessment Test (SweSAT): Development, results and experiences* (EM No. 49). Umeå, Sweden: Umeå University, Department of Educational Measurement.

Sungworn, N. (2009). *An investigation of using collateral information to reduce equating biases of the post-stratification equating method*, PhD thesis. Michigan State University.

van der Linden, W. J. (Ed.). (2018). *Handbook of item response theory: Three volume set*. CRC Press.

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

Wallin, G., Häggström & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating? *Applied Psychological Measurement, 45*(7-8), 518-535. https://doi.org/10.1177%2F01466216211040486.

Wallin, G. & Wiberg, M. (2019). Propensity scores in kernel equating for non-equivalent groups. *Journal of Educational and Behavioral Statistics. 44*(4), 390-414. https://doi.org/10.3102/1076998619838226

Wallin, W. & Wiberg, M. (2020). Model selection for presmoothing of bivariate score distributions in kernel equating. In Wiberg, M., González, J., & Molenaar, D., Böckenholt, U., & Kim, S-J. (Eds.) (2020). Quantitative Psychology – 84th Annual Meeting of the psychometric society, Santiago, Chile, 2019, New York: Springer. 97-105.

Wallin, G. & Wiberg, M. (2023). Model misspecification and robustness of test score equating using propensity scores. *Journal of Educational and Behavioral Statistics,* https://doi.org/10.3102/10769986231161575

Wallin, G. & Wiberg, M. (2024). Smoothing bivariate test score distributions - model selection targeting test score equating. *Journal of Educational and Behavioral Statistics, In press.*

Wallmark, J., Josefsson, M. & Wiberg, M. (2023). Efficiency analysis of item response theory kernel equating for mixed-format tests. *Applied Psychological Measurement. 47*(7-8), 496-512. https://doi.org/10.1177/01466216231209757

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

Wiberg, M. & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating, *Journal of Educational Measurement*, *53*(1), 106-125.

Wiberg, M., González, J., & von Davier (2025). *Generalized kernel equating with applications in R*, Boca Raton, FL: CRC Press.

Wiberg, M., Lyrén, P-E, & Lind Pantzare, A. (2021). Schools, Universities and Large-Scale Assessment Responses to COVID-19: The Swedish Example. *Education Sciences. 11*(175), 1-16.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.

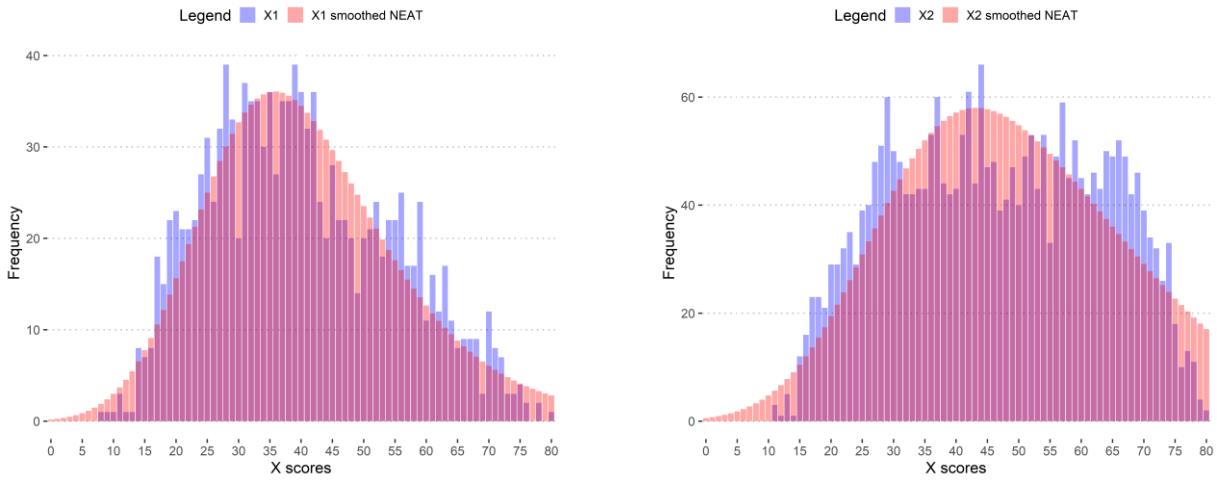
Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*(511), 910-922.

Appendix A

The comparison between the distributions of non-smoothed and smoothed with NEAT model Form X scores for scenarios 1 and 2 is presented in Figure A1. Smoothing results with other models are almost identical to the presented ones and thus are omitted.

Figure A1

The comparison between the distributions of non-smoothed and smoothed with NEAT model Form X scores for scenarios 1 and 2

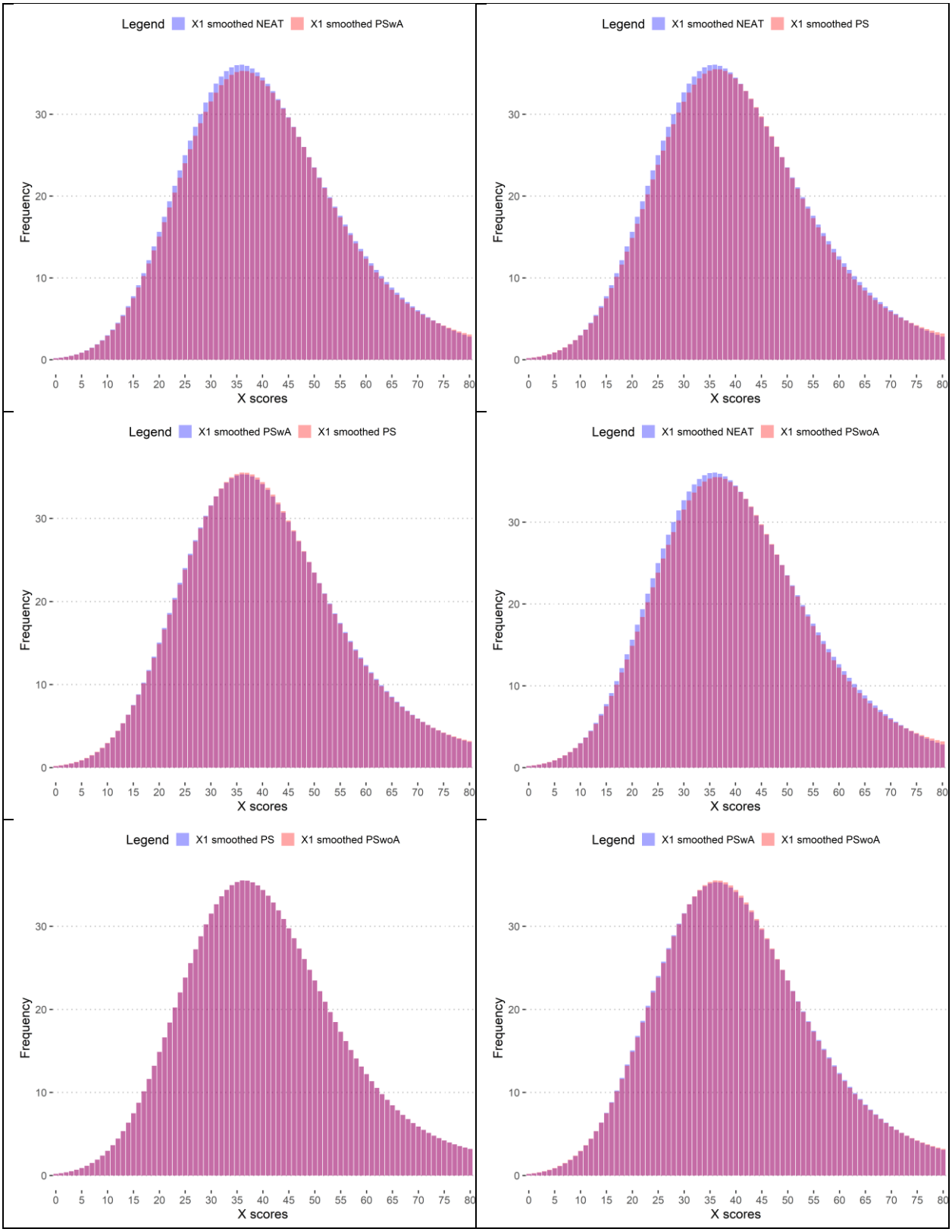


Scenario 1 Scenario 2

Figure A2 shows the comparison of smoothed distributions between different models used in the study. Only the results for Scenario 1 are presented here, as the results for Scenario 2 are almost identical.

Figure A2

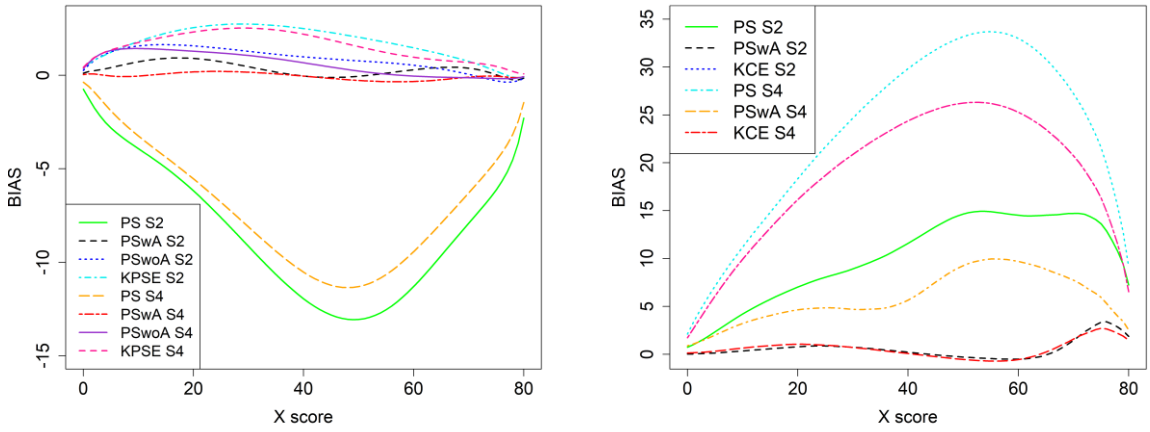
The comparison between the methods for Scenario 1



Appendix B

Figure B1

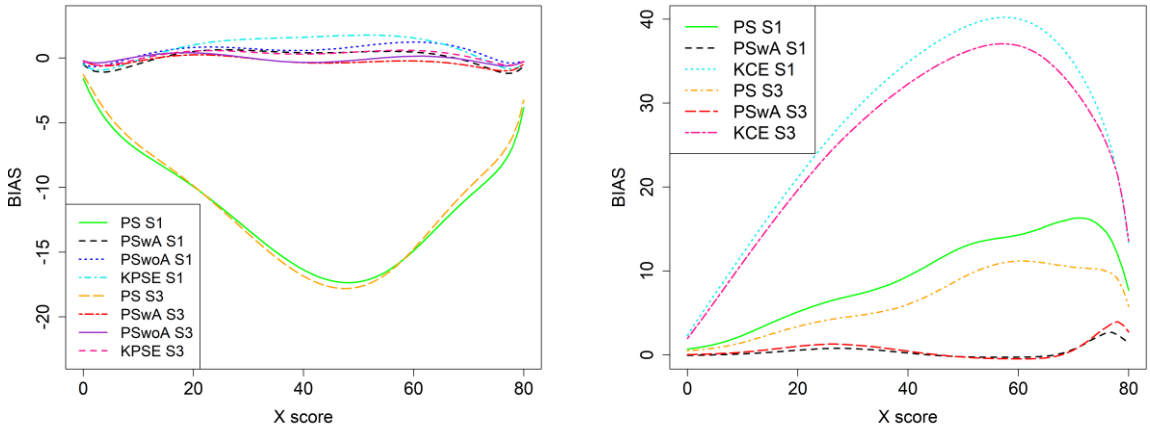
Bias for the baseline case when correlation between covariates was moderate and the size of an anchor test form was either 20 items (S2) or 40 items(S4)



a) b)

Figure B2

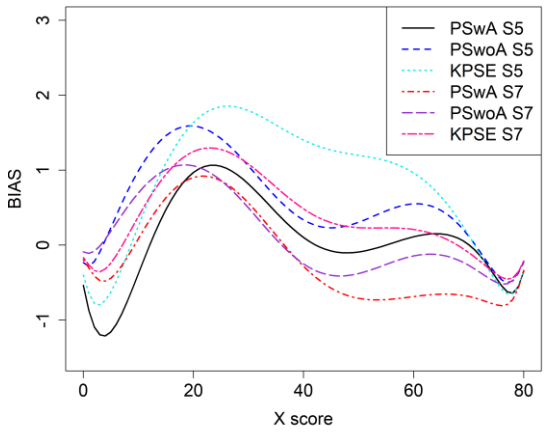
Bias for the baseline case when correlation between covariates was weak and the size of an anchor test form was either 20 items (S1) or 40 items(S3)



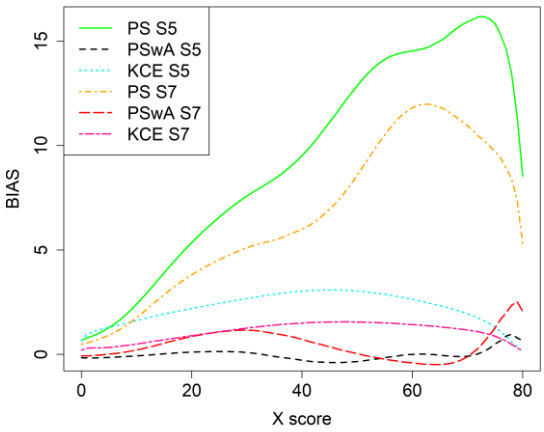
a) b)

Figure B3

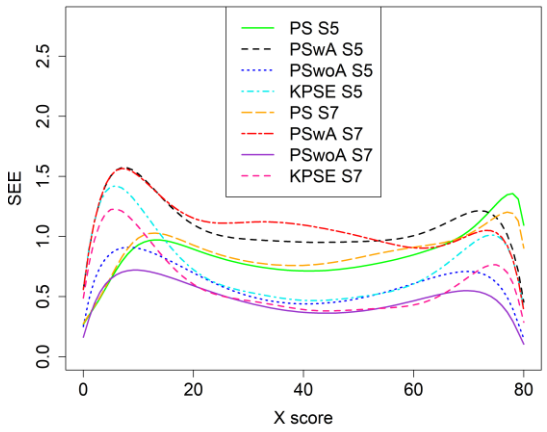
Bias (a and b) and SEE (c and d) for groups differing in ability when correlation between covariates was weak and the size of an anchor test form was either 20 items (S5) or 40 items(S7).



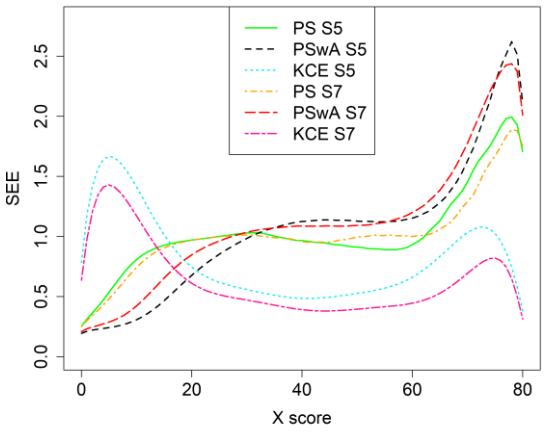
a)



b)



c)



d)