



Model Misspecification and Robustness of Observed-Score Test Equating Using Propensity Scores

Gabriel Wallin 

London School of Economics and Political Science
Umeå University

Marie Wiberg

Umeå University

This study explores the usefulness of covariates on equating test scores from nonequivalent test groups. The covariates are captured by an estimated propensity score, which is used as a proxy for latent ability to balance the test groups. The objective is to assess the sensitivity of the equated scores to various misspecifications in the propensity score model. The study assumes a parametric form of the propensity score and evaluates the effects of various misspecification scenarios on equating error. The results, based on both simulated and real testing data, show that (1) omitting an important covariate leads to biased estimates of the equated scores, (2) misspecifying a nonlinear relationship between the covariates and test scores increases the equating standard error in the tails of the score distributions, and (3) the equating estimators are robust against omitting a second-order term as well as using an incorrect link function in the propensity score estimation model. The findings demonstrate that auxiliary information is beneficial for test score equating in complex settings. However, it also sheds light on the challenge of making fair comparisons between nonequivalent test groups in the absence of common items. The study identifies scenarios, where equating performance is acceptable and problematic, provides practical guidelines, and identifies areas for further investigation.

Keywords: test score equating; kernel equating; nonequivalent groups; propensity scores; model misspecification

1. Introduction

Test score equating is a crucial statistical tool that enables the comparison of test scores from different test forms and ensures fairness in assessments

(González & Wiberg, 2017). When equating scores from nonequivalent test groups, it is essential to account for differences in both the ability levels of the test groups and the difficulty of the test forms. To make the scores comparable, any differences in ability and difficulty must be separated, so that the scores are only adjusted for differences in difficulty. For this purpose, testing programs generally apply either an assumption of common test-takers or the use of common items. The former assumes that the test groups to be equated are random samples from the same underlying population, whereas the latter views the groups as samples from different populations. In the latter case, a subset of common items is used to adjust for the differences in ability between the test groups. These common items are often referred to as anchor items and the belonging data collection design is known as the Nonequivalent Groups With Anchor Test (NEAT) design (von Davier et al., 2004b). However, not all testing programs have common items available but still need to adjust for ability imbalances. Examples of such tests are the Invalsi test (Invalsi, 2013), the Armed Services Vocational Aptitude Battery (Quenette et al., 2006), and the Swedish Scholastic Aptitude Test (SweSAT; Stage & Ögren, 2004) up until recently. If the ability imbalances are ignored, the equated test scores will be biased, which can have severe consequences in high-stakes testing scenarios.

One way of applying a nonequivalent groups design without anchor items is to use background information about the test takers in the form of measured covariates (Wiberg & Bränberg, 2015). There are several ways that covariates can be utilized within equating. Kolen (1990), Cook et al. (1990), and Wright and Dorans (1993) used covariates to balance the test groups before equating the test forms, Liou et al. (2001) applied covariates in a similar fashion to anchor items, Bränberg and Wiberg (2011) incorporated covariates in linear equating, and Hsu et al. (2002) used covariates within item response theory (IRT) true-score equating. However, as the covariate vector grows, controlling for the covariates quickly becomes very difficult. For example, conditioning on four categorical covariates, each with four categories, yields 256 possible outcomes. The matrix of all possible combinations of test scores and covariate realizations would therefore have an inflated number of empty cells. To overcome this problem, the test-takers can be compared on their propensity score instead, which is a scalar function of the covariates.

Livingston et al. (1990) was the first to propose the use of covariates within a propensity score for equating. More recently, Moses et al. (2010) explored the use of two anchor tests within a propensity score, Powers (2010) applied chained equating (CE) frequency estimation, IRT true score, and observed-score equating using propensity scores, Haberman (2015) used propensity scores to create pseudo-equivalent groups from nonequivalent groups, and Longford (2015) used it as a tool for matching before equating. Wallin and Wiberg (2019) were the first to propose propensity scores for both a poststratification equating (PSE) and CE estimator within the kernel equating framework (von Davier et al., 2004b). Their

results showed that a similar level of precision and accuracy compared to the NEAT design could be achieved. However, their results were based on the assumption that the propensity score was known. Since this will never be the case in any real testing situation, it is of great importance to assess the sensitivity of violations of this assumption. Thus, the aim is to study the functional form of the propensity score through which the covariates are conditioned on and investigate how sensitive the equated scores are to model misspecification of the estimated propensity score using both real and simulated data.

Propensity score model misspecification has previously been studied within the field of causal inference. Drake (1993) showed that a substantial bias was introduced when estimating the average treatment effect if a confounding covariate was omitted in the propensity score estimation model. Dehejia and Wahba (1999) had similar findings but also noted that causal estimates were not sensitive to the specification of the functional form of the propensity score, once all important covariates had been included. This has been shown in more recent studies as well, where Waernbaum (2010, 2012) showed that the average treatment effect can be unbiasedly estimated using propensity scores even when, for example, the link function is misspecified or when failing to include higher order terms of the covariates. There were furthermore situations with no efficiency loss, and one of the key components to obtain such results was that the true propensity score was a function of the misspecified model.

There are currently no existing studies on propensity score model misspecification in the equating context. This is critical to examine since the equating results often are used for decision-making on an individual level (e.g., admission decisions to universities) and for educational policy making. The current study therefore investigates the sensitivity of the equating function for model misspecification of the propensity score. Assuming a parametric model for the propensity score, three misspecifications are considered, inspired by the studies of Waernbaum (2010) and Waernbaum (2012): (1) misspecifying the link function, (2) excluding an important (true confounder) covariate, and (3) excluding a higher order moment of a confounding covariate. Each misspecification will be evaluated in terms of the equating function precision and accuracy to determine how critical they are.

The structure of this article is as follows. The kernel equating framework is introduced in Section 2, followed by an introduction to propensity scores in Section 3. Section 4 includes an empirical illustration, and Section 5 presents a simulation study. This article is concluded with a discussion of the results together with some practical guidelines.

2. Kernel Equating

We denote the new test form by X and the old test form by Y and their respective scores by X and Y . The realizations of X and Y are denoted x_j ,

$j = 1, \dots, J$, and y_k , $k = 1, \dots, K$. The test-takers receiving test form X are viewed as a random sample from population P , and the test-takers receiving test form Y as a random sample from population Q . With randomly sampled groups, the score variables X and Y are considered being random variables with sample spaces \mathcal{X} and \mathcal{Y} . An equating function thus maps the test scores from \mathcal{X} to \mathcal{Y} . However, not all such functions are considered an equating function. See Kolen and Brennan (2014) for a list of requirements.

Consider the random variable $\xi = F(X)$, which is well-known to follow a uniform distribution on the interval $[0, 1]$, given that F is a continuous and strictly increasing cumulative distribution function (CDF). It is consequently true that $V = G^{-1}(\xi)$ exactly follows the distribution given by G , as long as G has a properly defined inverse. The equipercentile function (Braun & Holland, 1982) is undoubtedly the most common equating function and uses this simple relationship between distributions of continuous random variables. With G_T and F_T denoting the CDFs of Y and X on the target population T for the equating parameter, the equipercentile equating function is defined as

$$\phi(x) = G_T^{-1}(F_T(x)). \quad (1)$$

The equipercentile function thus matches all of the moments of Y by matching the scores from X and Y that are at the same quantile of their respective distributions, that is

$$F_T(x) = u = G_T(y), u \in (0, 1). \quad (2)$$

However, since most test scores are discrete, their CDFs are not continuous but step functions. Hence, for any value $u \in (0, 1)$, it is rarely the case that there are two scores x and y that satisfy Equation 2. All test score equating methods that utilize the equipercentile function in (1) therefore need to resolve this issue.

Since kernel equating (Holland & Thayer, 1989; von Davier et al., 2004b) generalizes many of the most common and modern equating approaches, we present our theory in terms of this framework although the proposed method is applicable for example traditional equipercentile and linear equating as well. This framework consists of five steps: (1) fitting a regression model (typically a log-linear model) to the empirical score distributions, (2) estimating the test score probabilities on the target population based on the estimated model in Step 1 and given the data collection design, (3) making continuous approximations to the estimated discrete score distributions from Step 2, (4) equating the test scores using the equipercentile function, and (5) evaluating the estimated equating function (González & Wiberg, 2017; von Davier et al., 2004b). From Equation 1, it is clear that in order to estimate $\phi(\cdot)$, we need estimators of F_T and G_T . Kernel equating first uses the maximum likelihood estimates of the test score probabilities $r_j = P(X = x_j)$ and $s_k = P(Y = y_k)$ and then makes continuous approximations of these distributions using kernel functions. It is thus a

semiparametric method of estimating the equating function $\phi(\cdot)$. For this purpose, we define the joint distribution of (X, A) and (Y, A) , where A denotes a proxy variable for the latent ability that the test is constructed to measure. Typically, A represents an anchor test score, but as will be presented in the next section, we will instead consider a set of covariates that are gathered in a propensity score. Let $\mathbf{P} = \{p_{jl}\}_{J \times L}$, where $p_{jl} = \Pr(X = x_j, A = a_l | P)$, $j = 1, \dots, J$ and $l = 1, \dots, L$. Letting $\mathbf{p}_l = (\mathbf{p}_{1l}, \dots, \mathbf{p}_{Jl})^T$, we vectorize the matrix \mathbf{P} , such that the vectors \mathbf{p}_l , $l = 1, \dots, L$, are stacked onto each other. We denote this by $v(\mathbf{P})$. For details, see von Davier et al. (2004b). It is common practice to fit a log-linear model to the data to reduce sampling variance, so we will assume that $v(\mathbf{P})$ can be described by a log-linear model with R number of free parameters:

$$\log(v(\mathbf{P})) = \boldsymbol{\alpha} + \mathbf{u} + \mathbf{B}^T \boldsymbol{\beta}, \quad (3)$$

where $\boldsymbol{\alpha}$ is a normalizing constant, \mathbf{u} is a known constant of length J that specifies the null model when $\boldsymbol{\beta} = \mathbf{0}$, $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_J)$ is a matrix of dimension $R \times J$ of known constants, and $\boldsymbol{\beta}$ is a R -dimensional vector of unknown parameters. Equivalent model assumption is made for $v(\mathbf{Q}) = (\mathbf{q}_1, \dots, \mathbf{q}_L)^T$, where $\mathbf{Q} = \{q_{jl}\}_{J \times L}$, $q_{jl} = \Pr(Y = y_j, A = a_l | Q)$, $j = 1, \dots, J$ and $l = 1, \dots, L$. The model parameters in (3) are estimated through maximum likelihood.

The next step is to estimate the score probabilities $\mathbf{r} = (r_1, \dots, r_J)^T$ and $\mathbf{s} = (s_1, \dots, s_K)^T$, where \mathbf{r} and \mathbf{s} are functions of $v(\mathbf{P})$ and $v(\mathbf{Q})$, respectively, and of a design function that depends on the choice of data-collection design and equating estimator. We will save the introduction of necessary assumptions for Section 3, where two propensity score-based estimators are presented. For now, we assume that legitimate estimators of \mathbf{r} and \mathbf{s} are available. We also choose to present the required quantities only for the X scores since the expressions for the Y scores are given by corresponding formulas.

Let the mean and variance of X be denoted by (μ_X, σ_X^2) , and let V denote a continuous random variable, such that $\mathbb{E}(V) = 0$ and $\mathbb{V}(V) = \sigma_V^2$. Lastly, let

$$a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_V^2 h_X^2},$$

where $h_X > 0$ denotes a smoothing parameter from here on referred to as the bandwidth. With the introduced notation, we define a new random variable \tilde{X} by constructing a linear combination of X , h_X , and V . This will serve as a continuous version of X :

$$\tilde{X} = a_X(X + h_X V) + (1 - a_X)\mu_X.$$

The random variable \tilde{X} is defined, such that $\mathbb{E}(\tilde{X}) = \mathbb{E}(X) = \mu_X$ and $\mathbb{V}(\tilde{X}) = \mathbb{V}(X) = \sigma_X^2$. To define the CDF of \tilde{X} , let $K_V(\cdot)$ denote the kernel function of V . It is then straight-forward to show that the CDF of \tilde{X} is equal to

$$F_{\tilde{X}}(x) = P(\tilde{X} \leq x) = \sum_j r_j K_V \left(\frac{x - a_X x_j - (1 - a_X) \mu_X}{a_X h_X} \right). \quad (4)$$

In most studies of kernel equating, the function $K_V(\cdot)$ is set to the standard normal CDF $\Phi(\cdot)$, although other choices have been suggested (Lee & von Davier, 2011). In the empirical illustration of this study, the Gaussian kernel function is used although the proposed estimators can be used together with any other proper kernel function. Since $F_{\tilde{X}}$ is a function of the quantities μ_X , σ_X , and a_X , which in turn all are a function of \mathbf{r} , every component needed to estimate the continuized score CDFs, except for the bandwidth h_X , is available after estimating \mathbf{r} .

The bandwidth h_X determines the degree of smoothness of the score distribution $F_{\tilde{X}}$ and is often selected by minimizing certain criterion function. The most commonly used criterion, first suggested in von Davier et al. (2004b), is given by

$$Q(h_X) = \sum_j (\hat{r}_j - \hat{f}_{\tilde{X}}(x; h_X))^2 + \kappa \sum_j A_j, \quad (5)$$

where $\hat{f}_{\tilde{X}}(x; h_X)$ is the density function of \tilde{X} for bandwidth h_X yielded by differentiating $F_{\tilde{X}}(x)$ in x , $A_j = 1$ if

$$[(\hat{f}'_{\tilde{X}}(x_j - \omega) > 0) \cap (\hat{f}'_{\tilde{X}}(x_j + \omega) < 0)] \cup [(\hat{f}'_{\tilde{X}}(x_j - \omega) < 0) \cap (\hat{f}'_{\tilde{X}}(x_j + \omega) > 0)],$$

and $A_j = 0$ otherwise (von Davier, 2013; Wallin et al., 2021). The weight κ could be chosen through, for example, cross-validation but is typically set to 1, and ω determines the neighborhood for which the criterion function penalizes a bandwidth that permits sign changes in f' . In this study, we use $\omega = 0.25$ as it has yielded densities that closely follows the raw score histograms. However, it has been shown that the equated scores are not sensitive to the choice of bandwidth among the methods that are currently available (Wallin et al., 2021). In both the empirical illustration and the simulation study, we therefore use the criterion function in (5).

Remark 1. As the bandwidth grows to infinity, the continuous score CDF $F_{\tilde{X}}(x) \approx \Phi\left(\frac{x - \mu_X}{\sigma_X}\right)$, which makes the KE estimator approach the linear equating function $\text{Lin}(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$. See von Davier et al. (2004b) for the proof. If the bandwidth is set large, for example, $h_X = 10\sigma_X$, the linear equating estimator can be closely approximated. If the bandwidth instead is set to something very small, $F_{\tilde{X}}$ is a close approximation of the step function F . The traditional percentile rank method, where $F_{\tilde{X}}$ and $G_{\tilde{Y}}$ are the piecewise linear functions, can thus also be closely approximated (von Davier et al., 2004b). These two results emphasize that KE comprises a family of equating methods that incorporates both of the traditional methods as special

TABLE 1.

The Nonequivalent Groups With Covariate (NEC) Design Summarized

NEC	X	Y	\mathbf{D}
P sample	✓		✓
Q sample		✓	✓

cases when the bandwidth is either very large (linear equating) or very small (percentile rank method).

With the estimated, continuized score distributions $\hat{F}_{\hat{X}}(x) = F_{\hat{X}}(x; \hat{\mathbf{r}})$ and $\hat{G}_{\hat{Y}}(y) = G_{\hat{Y}}(y; \hat{\mathbf{s}})$, the kernel equating estimator of the equipercetile function $\varphi(x)$ equals

$$\varphi(x; \hat{\mathbf{r}}, \hat{\mathbf{s}}) = G_{\hat{Y}}^{-1}(F_{\hat{X}}(x; \hat{\mathbf{r}}); \hat{\mathbf{s}}). \quad (6)$$

The asymptotic distribution of $\varphi(x; \hat{\mathbf{r}}, \hat{\mathbf{s}})$ is given by $\mathcal{N}(\varphi(x; \mathbf{r}, \mathbf{s}), \mathbf{J}_{\varphi_Y} \mathbf{J}_{\mathbf{DF}} \mathbf{C} \mathbf{C}' \mathbf{J}_{\mathbf{DF}}' \mathbf{J}_{\varphi_Y}')$, where \mathbf{J}_{φ_Y} denotes the Jacobian of the equating function, $\mathbf{J}_{\mathbf{DF}}$ denotes the Jacobian of the design function, and \mathbf{C} is the covariance matrix of the score distributions $v(\mathbf{P})$ and $v(\mathbf{Q})$. See Wallin and Wiberg (2019) for the specific formula of these quantities. The standard error of equating (SEE; von Davier et al., 2004b) is consequently given by

$$\|\mathbf{J}_{\varphi_Y} \mathbf{J}_{\mathbf{DF}} \mathbf{C}\|, \quad (7)$$

where $\|\cdot\|$ denotes the Euclidean norm.

3. Nonequivalent Groups With Covariate (NEC) Design

This section will clarify the viewpoint we take on the nonequivalent groups designs in test score equating, and the specific assumptions underlying the NEC design (Wiberg & Bränberg, 2015). The NEC design assumes that the group of test-takers being administered test form X are a random sample from population P , and the group of test-takers being administered test form Y are a random sample from population Q , where $P \neq Q$ and $X \neq Y$. Each test-taker thus has a recorded test score on only one of the test forms, but never both. Additional to the test score there is a vector of measured covariates $\mathbf{D} = (D_1, \dots, D_m)$ for all test-takers regardless of test form. The NEC design is summarized in Table 1.

The covariates in \mathbf{D} take a similar role to that of the anchor score in the NEAT designs, meaning that they are intended to adjust for any imbalance in ability between the test groups. All covariates confounding the relationship between the test form assignment mechanism and (X, Y) need to be controlled for. We denote the test form assignment by $Z = 1$ if a randomly chosen test-taker is administered test form X and $Z = 0$ if test form Y is administered.

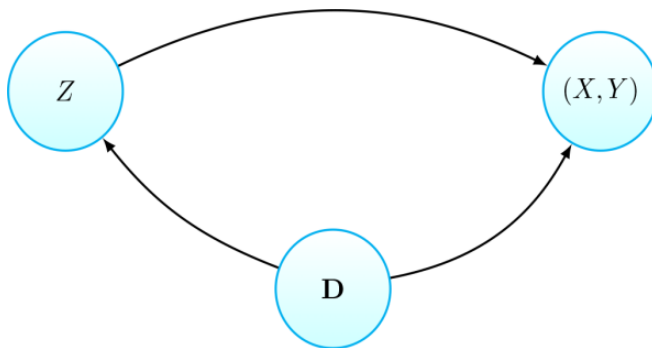


FIGURE 1. *The nonequivalent groups with covariates design.*

In Figure 1, the variables Z , \mathbf{D} , and (X, Y) are illustrated in a directed acyclic graph (DAG). In the DAG, the relationship between test form assignment and test score is confounded by the covariate vector \mathbf{D} . A proper equating procedure under the NEC design thus needs to control for such disturbance or else it will result in biased equated scores. In this sense, there is no difference with the use of anchor test scores A in NEAT design equating. Simply replace \mathbf{D} with A in the DAG, and it will graphically summarize the NEAT design. Just as the anchor test score A , the covariate vector \mathbf{D} thus is used as proxy for ability.

3.1. Propensity Scores

The basic idea of the NEC design is to replace the anchor test scores with the covariates and then to equate the test scores treating the covariate realizations as if they were in fact anchor scores. When using more than only a few covariates, the number of empty cells in the frequency table will grow large. There is thus a practical problem with the NEC design that is unrelated to the theoretical justification of the method. The curse of dimensionality is a well-known problem far beyond the equating literature, and a well-established method to handle this problem is by using a dimension-reducing function of the covariates called the propensity score. It reduces the dimension of covariate vector down to a scalar and is defined as $e(\mathbf{D}) = P(Z = 1|\mathbf{D})$.

The propensity score possesses the appealing property of being a balancing score (Rosenbaum & Rubin, 1983). This means that it is sufficient to control for $e(\mathbf{D})$ to balance the covariates between the test groups, if all confounders of the relationship between Z and (X, Y) are contained in \mathbf{D} . It is worth reminding that the variable we truly wish to control for is latent ability. It follows that the usefulness of balancing the test groups on the covariates is dependent on the quality of \mathbf{D} as a proxy variable for ability. Note that this is completely in line with the assumptions underlying NEAT-based equating using anchor scores.

As the propensity score is not known, it needs to be estimated. A common method is to use logistic regression, which will be used here. Following Rosenbaum and Rubin (1984) and Wallin and Wiberg (2019), the estimated propensity scores of the test-takers will thereafter be partitioned into strata based on the percentiles. The test-takers in each stratum are treated as homogeneous in terms of the latent ability, meaning that the equivalent groups design assumptions hold true within each stratum.

3.2. Equating Estimators Based on the Propensity Score

In the following, two propensity score-based equating estimators are derived and presented together with their underlying assumptions, following the estimators presented in Wallin and Wiberg (2019). Note that as these estimators were presented without much theoretical justification in the original paper by Wallin and Wiberg (2019), special attention is given to motivate them in this section.

3.2.1. PSE estimator. To define the PSE estimator, abbreviated PS-PSE, define the elements in \mathbf{r} and \mathbf{s} as

$$r_j = P(X = x_j|T) = wr_{Pj} + (1 - w)r_{Qj}, \quad (8)$$

and

$$s_k = P(Y = y_k|T) = ws_{Pk} + (1 - w)s_{Qk}, \quad (9)$$

where

$$\begin{aligned} r_{Pj} &= P(X = x_j|P) & r_{Qj} &= P(X = x_j|Q) \\ s_{Pk} &= P(Y = y_k|P) & s_{Qk} &= P(Y = y_k|Q) \end{aligned}.$$

The probabilities are defined on the target population T and populations P and Q . For PSE, this is a somewhat theoretical construct and described by the symbolic equation $T = wP + (1 - w)Q$, where w is a weight often set according to the relative sample sizes.

Remark 2. Every test score equating method needs to specify the target population T , which the equating function is defined for. Since one of the five requirements of any test score equating method is *population invariance*, the estimated equating function is independent of the subgroup of the population that is used to calculate it. By varying T , the fulfillment of this requirement can be empirically checked (Dorans & Holland, 2000). As is stated in von Davier et al. (2004b, p. 6), “. . . the use of a common target population is the way that we control for differential examinee ability in observed-score test equating.” Our view is therefore to consider P and Q as mutually exclusive and exhaustive strata of a larger mixture population T , and by varying the weight w , we have both in our empirical and simulation studies verified that the resulting equated scores have not changed in any practically important way. A discussion on the choice of w can be found in, for example, Brennan and Kolen (1987b), Angoff (1987), and Brennan and Kolen (1987a).

In Equations 8 and 9, the terms r_{Q_j} and s_{P_k} are not possible to calculate with data since the P sample has only been administrated test form X and the Q sample only test form Y . There is thus data missing by design, in the same sense as is thoroughly discussed in Sinharay and Holland (2010b). The following assumption is therefore needed to define the PS-PSE estimator. We follow the notation in Dawid (1979) and let $\perp\!\!\!\perp$ denote statistical independence.

Assumption 1: For the PS-PSE estimator, we assume that

$$(X, Y) \perp\!\!\!\perp Z | e(\mathbf{D}),$$

$$0 < e(\mathbf{D}) < 1.$$

for any $T = wP + (1 - w)Q$.

Note, for a dichotomous treatment (i.e., a pair of test forms to be equated), $P(X = x_j | Z = 1, e(\mathbf{D})) = P(X = x_j | Z = 0, e(\mathbf{D}))$ for all $j \in [1, J]$ and $P(Y = y_k | Z = 1, e(\mathbf{D})) = P(Y = y_k | Z = 0, e(\mathbf{D}))$ for all $k \in [1, K]$ if Assumption 1 is true. This is sometimes referred to as strong ignorability in the causal inference literature (Hernan & Robins, 2020).

The first part of Assumption 1 means that the test scores are conditionally independent of the test form assignment by controlling for the propensity score. The test groups would thereby be only randomly different from each other, as in the equivalent groups design. The second part of Assumption 1 is to ensure that all test-takers have a nonzero probability of being assigned either test form. If the propensity score has been stratified into L strata, such that the test groups are balanced on \mathbf{D} in each stratum, estimators of the missing-data quantities in Equations 8 and 9 can be identified under Assumption 1. To that end, let the stratified propensity score be denoted $M \in \{1, \dots, L\}$ with realizations denoted m .

Under Assumption 1, we furthermore assume that:

$$\Pr(X = x_j | M = m; Q) = \Pr(X = x_j | M = m; P), \quad (10)$$

and

$$\Pr(Y = y_k | M = m; P) = \Pr(Y = y_k | M = m; Q). \quad (11)$$

Equation 10 states that the probability of test score x_j is the same in populations P and Q conditional on the observed, stratified propensity score $M = m$. The corresponding probability statement is true for the Y scores in Equation 11. Estimators of r_{Q_j} and s_{P_k} are now possible to define. In the following, such estimators are defined and justified.

Proposition 1: Denote the (from log-linear models) estimated joint distributions of X and M in P , and of Y and M in Q , by

$$\hat{p}_{jl} = \Pr(X = x_j, M = m; P),$$

and

$$\hat{q}_{kl} = \Pr(Y = y_k, M = m; Q).$$

If Assumption 1 holds true and the propensity score has been stratified, such that the covariate distribution is balanced in the test groups, r_{Q_j} and s_{P_k} can be estimated by

$$\hat{r}_{Q_j} = \sum_l \left(\frac{\hat{p}_{jl}}{\sum_j \hat{p}_{jl}} \sum_k \hat{q}_{kl} \right),$$

and

$$\hat{s}_{P_k} = \sum_l \left(\frac{\hat{q}_{kl}}{\sum_k \hat{q}_{kl}} \sum_j \hat{p}_{jl} \right).$$

The proof of Proposition 1 is found in Online Appendix A.

Lastly, plug the estimated test score probability $\hat{\mathbf{r}}$ into Equation 4, and do the corresponding for $\hat{\mathbf{s}}$, and functionally compose the equating estimator as

$$\varphi(x; \hat{\mathbf{r}}, \hat{\mathbf{s}})_{\text{PSE}} = G_Y^{-1}(F_X(x; \hat{\mathbf{r}}); \hat{\mathbf{s}}). \quad (12)$$

3.2.2. CE estimator. Even though conditioning on the propensity score, as has been outlined for the PS-PSE estimator, is the traditional way of removing dependencies between the outcome and the treatment, CE methods have a long-standing tradition within test score equating. Several studies have showed that the result of linking, or *chaining*, together a sequence of equating functions is often very similar and sometimes even better than that of the competing PSE estimator (Sinharay & Holland, 2010a, 2010b; Wallin & Wiberg, 2019). In fact, when equating with an anchor, the PSE and CE coincide if the anchor score distribution in P and Q is equal (von Davier et al., 2004a). Thus, for the second equating estimator considered, abbreviated PS-CE, let

$$\begin{aligned} \mathbf{r}_P &= (r_{P1}, \dots, r_{PJ})^T & \mathbf{s}_Q &= (s_{Q1}, \dots, s_{QK})^T \\ \mathbf{t}_P &= (t_{P1}, \dots, t_{PJ})^T & \mathbf{t}_Q &= (t_{Q1}, \dots, t_{QK})^T, \end{aligned}$$

with $r_{Pj} = \Pr(X = x|P)$, $s_{Qk} = \Pr(Y = y|Q)$, $t_{Pj} = \Pr(M = m|P)$, and $t_{Qk} = \Pr(M = m|Q)$. Note that these are score probabilities for populations P and Q , respectively, and not for the mixture population T . The quantity t_{Pj} is to be understood as the probability of the random, stratified variable M being equal to the realization m in population P , and t_{Qk} interpreted analogously. We furthermore define the corresponding continuized score CDFs that are functions of the score probabilities:

$$\begin{aligned} F_{\tilde{X}_P}(x; \mathbf{r}_P) &= \Pr(\tilde{X} \leq x|P) \\ G_{\tilde{Y}_Q}(y; \mathbf{s}_Q) &= \Pr(\tilde{Y} \leq y|Q) \\ H_{\tilde{e}_P}(m; \mathbf{t}_P) &= \Pr(M \leq m|P) \\ H_{\tilde{e}_Q}(m; \mathbf{t}_Q) &= \Pr(M \leq m|Q) \end{aligned},$$

where $F_{\tilde{X}_P}$, $G_{\tilde{Y}_Q}$, $H_{\tilde{e}_P}$, and $H_{\tilde{e}_Q}$ denote CDFs that have been continuized in similar fashion as in Equation 4.

The PS-CE estimator is dependent on the linking of distributions between populations P and Q . There is an underlying assumption that there is a link between the X scores and the propensity scores in population P and a link between the propensity scores and the Y scores in population Q . This is specified in Assumption 2.

Assumption 2: *For the PS-CE estimator, we assume that*

$$\begin{aligned} H_{e_P}^{-1}(F_{\tilde{X}_P}(x)) &= H_T^{-1}(F_T(x)) \\ G_{\tilde{Y}_Q}^{-1}(H_{e_Q}(m)) &= G_T^{-1}(H_T(m)), \end{aligned} \tag{13}$$

for any target distribution of the form $T = wP + (1 - w)Q$.

Assumption 2 is to be understood as a statement regarding population invariance of the equipercetile function linking X to $e(\mathbf{D})$ on P and of the equipercetile function linking $e(\mathbf{D})$ to Y on Q .

Proposition 2: *The PS-CE estimator is given by linking the functions in Equation 13 together in a chain:*

$$\varphi(x; \hat{\mathbf{r}}_P, \hat{\mathbf{t}}_P, \hat{\mathbf{t}}_Q, \hat{\mathbf{s}}_Q)_{CE} = G_{\tilde{Y}_Q}^{-1}(H_{\tilde{e}_Q}(H_{\tilde{e}_P}^{-1}(F_{\tilde{X}_P}(x; \hat{\mathbf{r}}_P); \hat{\mathbf{t}}_P); \hat{\mathbf{t}}_Q); \hat{\mathbf{s}}_Q).$$

The proof of Proposition 2 is found in Online Appendix B.

4. A Motivating Example Using Empirical Data

As a motivating example, two test administrations of the SweSAT are analyzed. The SweSAT is used in the selection process for Swedish university programs and consists of a verbal and quantitative section. These sections in turn consist of 80 items each and are equated separately. Only recently, the SweSAT started including anchor items. Prior to this, covariates were used in a matching procedure when the test forms were equated (Wiberg & Bränberg, 2015). In this empirical study, both the PS-PSE and PS-CE estimators will be used to equate the quantitative sections from two SweSAT test administrations from the past decade.

4.1. Data and PS Models

The score distributions of the analyzed test forms are shown in Figure 2. As seen, the Y score distribution (the old test form) is slightly skewed and shifted to the left of the X score distribution. The empirical distributions suggest that either the X test group is on average more capable compared to the Y test group or that the X test form is easier, or a combination of both. In addition to the test scores, each test-taker has a set of covariates recorded. Based on previous studies

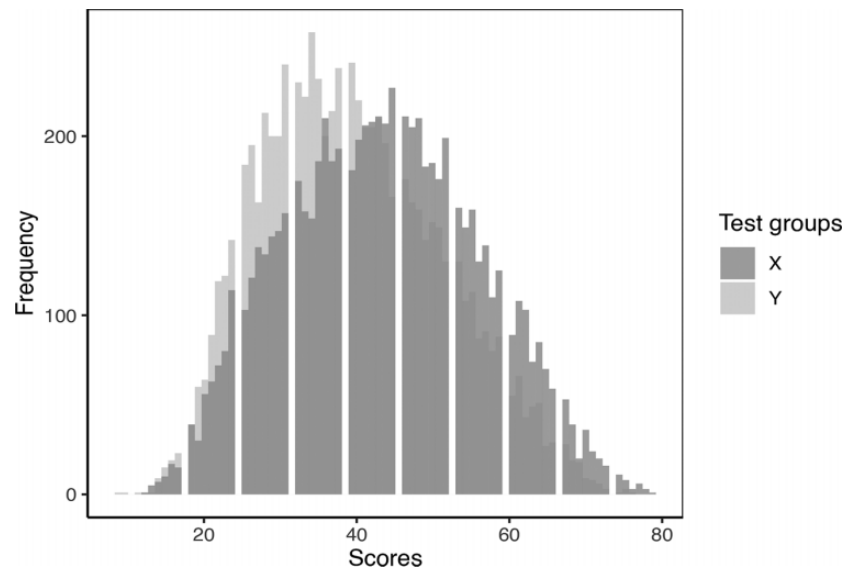


FIGURE 2. The score distributions of the X and Y scores, where X represents the new test form and Y the old test form.

TABLE 2.
Summary Statistics of the Variables Used in the Empirical Illustration

	Verb	Age	Gender	X	Y
Correlation to Y	0.48	−0.14	0.26	0.27	1
Correlation to X	0.52	−0.13	0.28	1	0.27
Mean	43.91 (39.35)	1 (1)	0.42 (0.53)	43.32	39.34
Standard deviation	12.08 (11.56)	2 (2)	0.49 (0.50)	12.65	11.80

Note. Verb refers to the verbal test score. Values within parentheses refer to form Y. The correlations for the variables Age and Gender are the Spearman and point-biserial correlations, respectively. For Age, the last two rows present the median and quartile deviation.

(Altıntaş & Wallin, 2021; Bränberg et al., 1990; Wallin & Wiberg, 2019) and on availability, the covariates used in this study are gender, age, and the test score from the verbal section as these have shown to correlate with the quantitative score. In Table 2, summary statistics are being presented for the variables of the empirical study. The variable Age is reported in five categories: It equals 1 if an individual’s age is within [0, 20], it equals 2 if the age is within [21 – 24], it equals 3 if the age is within [25 – 29], it equals 4 if the age is within [30 – 39], and it equals 5 if the age is 40 or older. Note that at the time for these test administrations, there was no age restriction for individuals taking the test.

TABLE 3.
The Parametrization of the Candidate Propensity Score Models

Covariate Combinations	
1. D_1, D_2 , and D_3	8. D_1 and D_3
2. D_1, D_2 , and D_3 with probit	9. D_2 and D_3
3. D_1 and D_2	10. D_2, D_2^2 , and D_3
4. D_1, D_1^2, D_2 , and D_3	11. D_1, D_1^2 , and D_2
5. D_1	12. D_1, D_2, D_2^2 , and D_3
6. D_2	13. D_1, D_1^2, D_2, D_2^2 , and D_3
7. D_3	

Since there is no known true propensity score model, a number of candidate models are set up for both the PS-PSE and PS-CE equating estimators. Let D_1 denote the test score on the verbal section of the test, D_2 denote age, and D_3 denote gender. The candidate propensity score models are estimated using logistic regression with a logit link except when indicated. We consider most of the possible combinations of covariates and factors, resulting in 13 models, which are shown in Table 3.

Hence, in total, there will be 26 equating estimators considered, 13 for the PS-PSE estimator and 13 for the PS-CE estimator. The equated scores and the SEEs of each estimator will be analyzed to determine the extent to which they vary with changes in the propensity score model's parameterization. The difference that matters (Dorans & Feigenbaum, 1994), defined to be larger than half a raw score point, will also be investigated. Goodness-of-fit measures like the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (BIC; Schwarz, 1978) are not suitable for evaluating the propensity score models, since their parameter estimates are not the priority but rather the achieved covariate balance between the test groups (Augurzy & Schmidt, 2001; Stuart, 2010). The absolute standardized mean difference (ASMD; Austin, 2008) will therefore be used to evaluate the level of achieved covariate balance:

$$\text{ASMD} = \left| \frac{\mu_D^{(T)} - \mu_D^{(C)}}{\sqrt{\frac{\sigma_D^{2(T)} + \sigma_D^{2(C)}}{2}}} \right|, \quad (14)$$

where $\mu_D^{(T)}$ and $\mu_D^{(C)}$ denote the means of the treatment (test form X) and control (test form Y) group for covariate D , respectively, and $\sigma_D^{2(T)}$ and $\sigma_D^{2(C)}$ denote their respective variances. There exist no general threshold for the ASMD, but a value above 0.10 is considered to indicate covariate imbalance (Austin, 2008). Once a proper stratification has been achieved, bivariate log-linear models of the test scores and the stratified, estimated propensity score according to (3) are fit to

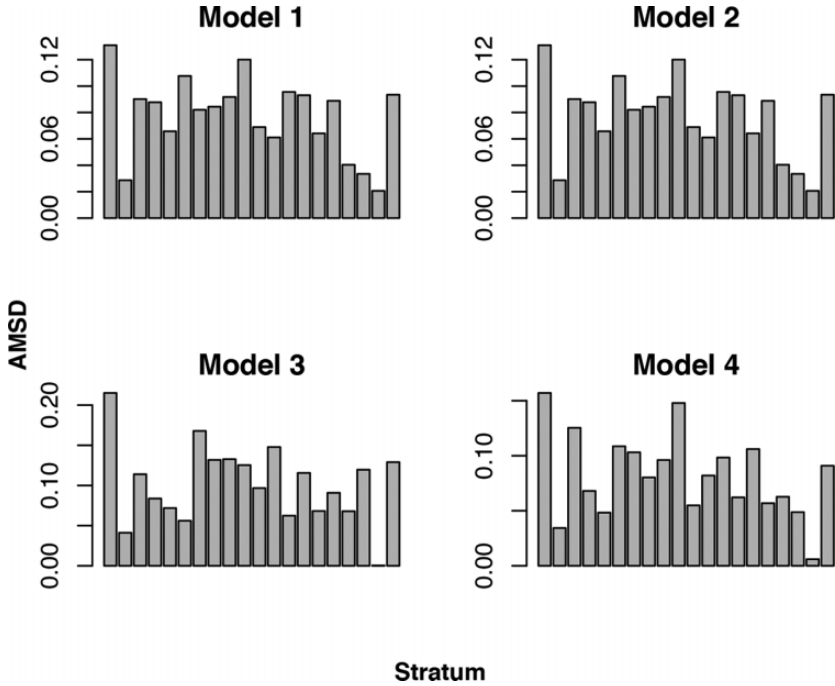


FIGURE 3. The absolute standardized mean difference between the treatment group (test form Y) and the control group (test form X) for the covariate Verb, for each of the 20 strata of the four candidate propensity score models.

the empirical data. The BIC is used to choose parametrization of the log-linear models since it has been proven to have a high selection accuracy for bivariate smoothing (Moses & Holland, 2010). All analyses are made using R (R Core Team, 2021) and the R package **kequate** (Andersson et al., 2013).

In Figure 3, the ASMDs between the treatment (test form X) and control (test form Y) group for the covariate Verb are displayed. The ASMD was calculated within each stratum of the propensity score. To determine the number of strata, a sequence of potential stratifications was set up. For each possible number of strata, the ASMD was calculated within each stratum. The stratification that produced a low ASMD for all strata was chosen, resulting in 20 strata for this particular dataset. In Figure 3, the first two models show the best performance in terms of ASMD since most of the strata have an ASMD below 0.1. This could be compared with the ASMD if not controlling for the propensity score, which equals 0.386. Stratifying on the propensity score has thus successfully brought the test groups substantially closer in terms of their covariate distribution. The corresponding plots were examined for the covariates Gender and Age as well,

TABLE 4.

The Estimated Coefficients, With Standard Errors in Parenthesis, of the Four Bivariate Log-Linear Models Fit Considered for the X Data

	Candidate Models			
	(1)	(2)	(3)	(4)
X	0.399*** (0.120)	0.399*** (0.120)	1.114*** (0.122)	1.163*** (0.124)
X^2	-0.077** (0.032)	-0.077** (0.032)	-0.087*** (0.032)	-0.085*** (0.032)
X^3	0.007** (0.004)	0.007** (0.004)	0.008** (0.004)	0.007** (0.004)
X^4	-0.0003* (0.0002)	-0.0003* (0.0002)	-0.0003* (0.0002)	-0.0003* (0.0002)
X^5	0.00001* (0.00000)	0.00001* (0.00000)	0.00001* (0.00000)	0.00001* (0.00000)
M	-0.001* (0.0005)	0.280*** (0.005)	0.749*** (0.019)	0.830*** (0.041)
M^2		-0.004*** (0.0001)	-0.010*** (0.0004)	-0.011*** (0.001)
M^3			0.00004*** (0.00000)	0.00005*** (0.00001)
$X : M$			-0.013*** (0.0003)	-0.016*** (0.002)
$X^2 : M$				0.00004** (0.00002)
Constant	0.531*** (0.150)	-4.023*** (0.176)	-15.762*** (0.384)	-17.055*** (0.699)
Observations	1,620	1,620	1,620	1,620
Bayesian information criterion	15,388.42	7,963.714	4,892.262	4,894.582

* $p < .1$. ** $p < .05$. *** $p < .01$.

where similar patterns were observed. It could thus be suspected that these models will lead to low equating error, given that the covariates Verb, Gender, and Age are at all associated with the latent ability.

In the next stage, bivariate log-linear models are fit to the observed test scores and the stratified propensity scores. A set of candidate models are considered and evaluated in terms of their BIC. In Tables 4 and 5, the estimated coefficients together with their corresponding standard errors, p values, and the BIC are presented for four candidate models. The notation $X : M$ refers to an included interaction term between X and Y . We decided to use the third model for both the

TABLE 5.

The Estimated Coefficients, With Standard Errors in Parenthesis, of the Four Bivariate Log-Linear Models Fit Considered for the Y Data

	Dependent Variable			
	Frequency			
	(1)	(2)	(3)	(4)
Y	−0.256*** (0.092)	−0.196** (0.093)	0.374*** (0.094)	0.362*** (0.096)
Y^2	0.051* (0.027)	0.036 (0.027)	0.029 (0.027)	0.029 (0.027)
Y^3	−0.005 (0.003)	−0.004 (0.003)	−0.004 (0.003)	−0.004 (0.003)
Y^4	0.0003 (0.0002)	0.0002 (0.0002)	0.0002 (0.0002)	0.0002 (0.0002)
Y^5	−0.00000 (0.00000)	−0.00000 (0.00000)	−0.00000 (0.00000)	−0.00000 (0.00000)
M	0.006*** (0.001)	0.265*** (0.005)	0.564*** (0.018)	0.548*** (0.032)
M^2		−0.003*** (0.0001)	−0.007*** (0.0004)	−0.006*** (0.001)
M^3			0.00002*** (0.00000)	0.00002*** (0.00000)
$Y : M$			−0.010*** (0.0003)	−0.010*** (0.001)
$Y^2 : M$				−0.00001 (0.00001)
Constant	1.949*** (0.101)	−2.700*** (0.141)	−10.566*** (0.326)	−10.307*** (0.544)
Observations	1,620	1,620	1,620	1,620
Bayesian information criterion	13,804.4	7,534.71	5,131.271	5,138.315

* $p < .1$. ** $p < .05$. *** $p < .01$.

X and Y scores, as it showed the best fit in terms of the BIC. We thereafter continuized the score distributions by applying a Gaussian kernel to the distribution approximation in Equation 4 and select the degree of smoothness using the criterion in Equation 5.

4.2. Results

To illustrate the general trend among the estimators, we display the results of the equating estimators using propensity score models 1–4 in Figure 4.

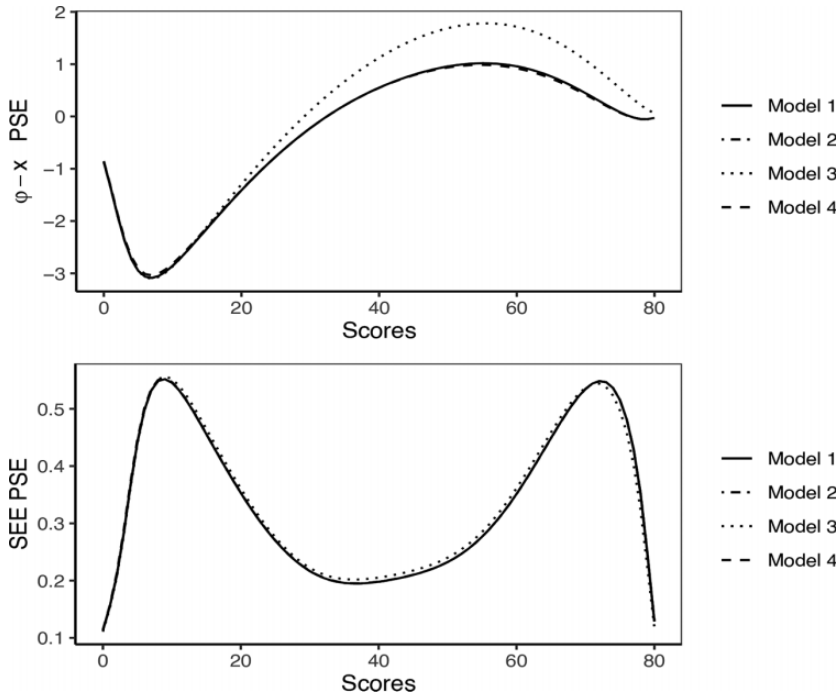


FIGURE 4. The equated scores and standard error of equating of the PS-PSE estimator, using Models 1–4 for the propensity score estimation.

Propensity score model number 3, which does not include the covariate Gender, deviates clearly. For the upper score scale, Model 3 has a score difference to the other estimators that clearly matters. Since gender has been established as an important covariate when analyzing the SweSAT (Bränberg et al., 1990) and with a fairly strong correlation with the test scores, it comes as no surprise that the equated scores are affected when gender is excluded. Far less important is the choice of link function, or whether or not a second-order term is included, for this dataset. On the other hand, the SEEs of all estimators are more or less similar along the whole score scale.

In Figure 5, the equated scores (upper part) are shown for the four PS-CE estimators, together with SEE (lower part). The pattern from the PSE estimators is evident here as well, with clear deviations for the model that fails to include gender in the propensity score model and with a negligible difference in terms of SEE. We also notice a distinct difference between the equated scores produced by the PSE-based estimators in Figure 4 and the CE-based estimators in Figure 5. In the online supplements, the estimated equating functions resulting from all 13 propensity score models are given.

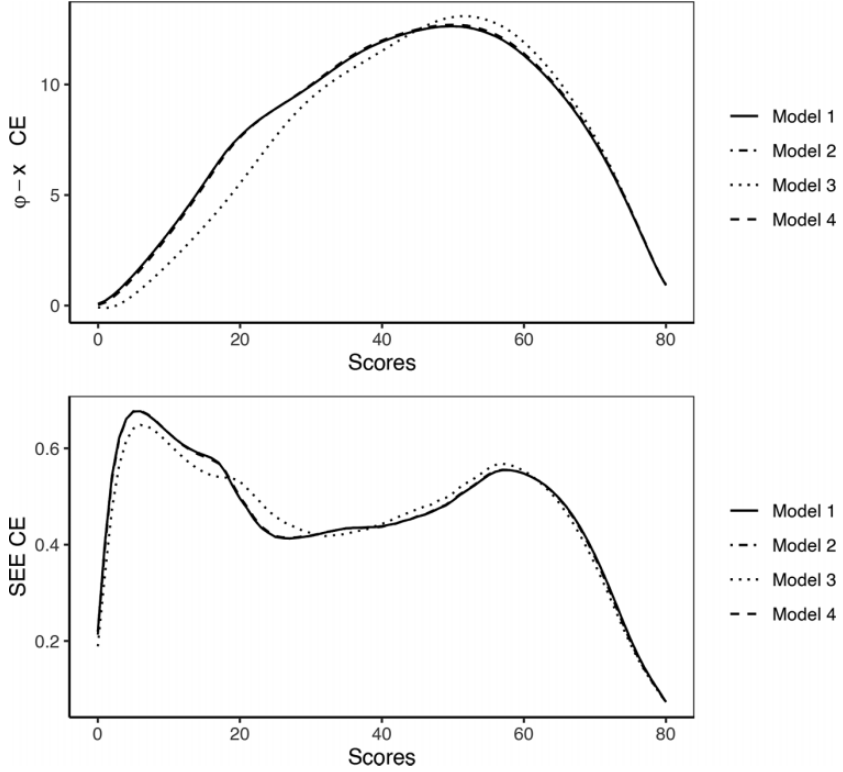


FIGURE 5. The equated scores and standard error of equating of the PS-CE estimator, using Models 1–4 for the propensity score estimation.

5. Simulation Study

For the empirical illustration, the results suggested that a critical component when using propensity scores to equate test scores is to include all important covariates in the propensity score estimation model. The equated scores were less sensitive to the choice of link function and the inclusion of higher order polynomials. Since it is not possible to generalize these results, the robustness of the PS-PSE and PS-CE estimators to misspecifications of the propensity score model is evaluated in a simulation study. We assume that the propensity score is described by a parametric model and consider two different simulation designs. Both designs are inspired by the simulation study in Wallin and Wiberg (2019) but with propensity score model misspecifications added. The misspecifications considered are (1) using the wrong link function, (2) leaving out a covariate, and (3) leaving out higher order terms. The simulation designs follow closely the

studies typically seen in the causal inference literature, where potential outcomes under different treatment regimes are generated and the observed outcomes depend on the realization of the treatment variable, which in turn is a function of a covariate vector. Inspired by this and by trying to mimic the situation described in Figure 1, we generated covariates that both affected the test form assignment (through the propensity score) and the test scores, making them true confounders. Both potential test scores and observed test scores are generated, as explained in the simulation designs. The presented results are based on $n = 10,000$ simulated test-takers and 1,000 iteration, although sample sizes of $n = 1,000$ and $n = 5,000$ were considered as well. As the difference of those results to the ones based on $n = 10,000$ was negligible, they have been excluded but can be sent upon request. As in the empirical study, all calculations are carried out in R with the R package **kequate**.

5.1. Simulation Design A

For Design A, the data generating process (DGP) is as follows:

1. Generate the covariates $D_1, D_2 \sim \text{Uniform}(1, 5)$.
2. Generate $n = \{1,000, 5,000, 10,000\}$ Bernoulli trials to compose the treatment variable $Z \sim \text{Bernoulli}(e(\mathbf{D}))$, where

$$e(\mathbf{D}) = \frac{1}{1 + \exp(-0.36 + 1.25D_1 + 1.25D_2 - 0.35D_1^2 - 0.35D_2^2)}. \quad (15)$$

It follows that the test groups will be of approximately the same size.

3. The potential test scores on test form X are for all test-takers generated as

$$X = -6 + 4D_1 + 5D_2 + \epsilon_X,$$

and the potential test scores on test form Y are for all test-takers generated as

$$Y = -9 + 3D_1 + 6D_2 + \epsilon_Y.$$

Since the covariates in these expressions represent the ability differences between the groups, the ϵ terms represent the difficulty of the test forms, where $\epsilon_X \sim \mathcal{N}(2, 1.5)$ and $\epsilon_Y \sim \mathcal{N}(0, 1)$. The means and variances of the test scores are $\mathbb{E}[X] = 23$, $\mathbb{E}[Y] = 18$, $\mathbb{V}[X] \approx 56.92$, and $\mathbb{V}[Y] = 61$. With the data generated, the distributions of the covariates differ between the test groups.

4. The observed test score for each test-taker is defined as

$$U = ZX + (1 - Z)Y.$$

To generate an observed score U^* for each test-taker, we set $U^* = \min(U, 40)$, which is to be understood as the rounded value of whichever is the smaller of U and 40. Although no generated score was smaller than 0,

such score would have been truncated to 0. The score range is therefore set to $[0, 40]$.

5. The propensity score is estimated using logistic regression. Based on the percentiles, it is thereafter divided into 20 categories. The number of categories was chosen trying to reach a covariate balance between the test groups as measured by the ASMD. Four candidate models will be defined: one that is correctly specified according to Equation 15, one that uses a probit link function instead of the correct logit link, one that leaves out D_2 , and one that leaves out D_1^2 and D_2^2 .

5.2. Simulation Design B

For Design B, the DGP is as follows:

1. Generate the covariates $D_1, D_2 \sim \text{Uniform}(1, 5)$.
2. Generate $n = \{1,000, 5,000, 10,000\}$ Bernoulli trials to compose the treatment variable $Z \sim \text{Bernoulli}(e(\mathbf{D}))$, where

$$e(\mathbf{D}) = \frac{1}{1 + \exp(0.8 + 0.72D_1 + 0.72D_2 - 0.25D_1^2 - 0.25D_2^2)}. \quad (16)$$

It follows that the test groups will be of approximately the same size.

3. The scores on test form X are for all test-takers generated as

$$X = 9 + 1.25D_1 + 1.25D_2 + D_1^2 + D_2^2 + D_1D_2 + \epsilon_X,$$

and the scores on test form Y are for all test-takers generated as

$$Y = 7.5 + 1.25D_1 + 1.25D_2 + D_1^2 + D_2^2 + D_1D_2 + \epsilon_Y,$$

where $\epsilon_X \sim \mathcal{N}(0, 1)$ and $\epsilon_Y \sim \mathcal{N}(5, 1.5)$. Note that the covariates in this design have a nonlinear relationship with the test scores and that there is an interaction term included. The means and variances of the test scores are $\mathbb{E}[X] \approx 46.17$, $\mathbb{E}[Y] \approx 49.67$, $\mathbb{V}[X] \approx 129.79$, and $\mathbb{V}[Y] \approx 131.04$.

4. The observed test score for each test-taker is generated as

$$U = ZX + (1 - Z)Y.$$

To generate an observed score U^* for each test-taker, we set $U^* = \min(U, 90)$, which is to be understood as the rounded value of whichever is the smaller of U and 90. Although no generated score was smaller than 0, such score would have been truncated to 0 as in Design A. The score range is therefore set to $[0, 90]$.

5. As in Design A, the propensity score is estimated using logistic regression and thereafter divided into 20 categories, based on the absolute standardized mean difference. Four candidate models will be used: one that is correctly

specified according to Equation 16, one that uses a probit link function instead of the correct logit link, one that leaves out D_2 , and one that leaves out D_1^2 and D_2^2 .

Remark 3. The potential test score X is to be interpreted as the test score that a test-taker would have got if they had been administered the X test form, and Y is the potential test score if test form Y had been administered. In this way, every test-taker has a potential, but not observed, test score on both forms. The observed test score U reflects the test form actually administered to each of the test-takers. In addition, for both Designs A and B, each test-taker has an observed covariate vector \mathbf{D} and an estimated propensity score $\hat{e}(\mathbf{D})$. Also, a discrete version of the covariates was considered by splitting them into five equally spaced groups, inspired by the DGP in Wiberg and Bränberg (2015). The reason for doing this was to mimic testing programs where only categorized versions of the background information have been stored, such as prespecified age intervals instead of the actual ages of the test-takers. Lastly, it is important to note that it is possible to define a true equating function for both DGPs given above, since each test-taker has a potential test score on both test forms.

Remark 4. Note that in Design A, the covariates are associated with the log odds of the propensity score and the test scores in a linear way, whereas in Design B, that relationship involves both higher order terms and interactions. In this way, we are able to investigate whether there is any connection between model complexity, model misspecification, and sensitivity of the equated scores.

5.3. Evaluation Measures

The PS-PSE and PS-CE estimators are evaluated by calculating the bias and SE , as given in Wiberg and González (2016):

$$\text{Bias}(\hat{\varphi}(x_i)) = \frac{1}{1000} \sum_{g=1}^{1000} (\hat{\varphi}^{(g)}(x_i) - \varphi(x_i)),$$

and

$$\text{SE}(\hat{\varphi}(x_i))^2 = \frac{1}{1000} \sum_{g=1}^{1000} (\hat{\varphi}^{(g)}(x_i) - \bar{\varphi}(x_i))^2,$$

where

$$\bar{\varphi}(x_i) = \frac{1}{1000} \sum_{g=1}^{1000} \hat{\varphi}^{(g)}(x_i),$$

and $\hat{\varphi}^{(g)}(x_i)$ denote the estimated equating function evaluated at x_i for replicate g , $g = 1, \dots, 1,000$.

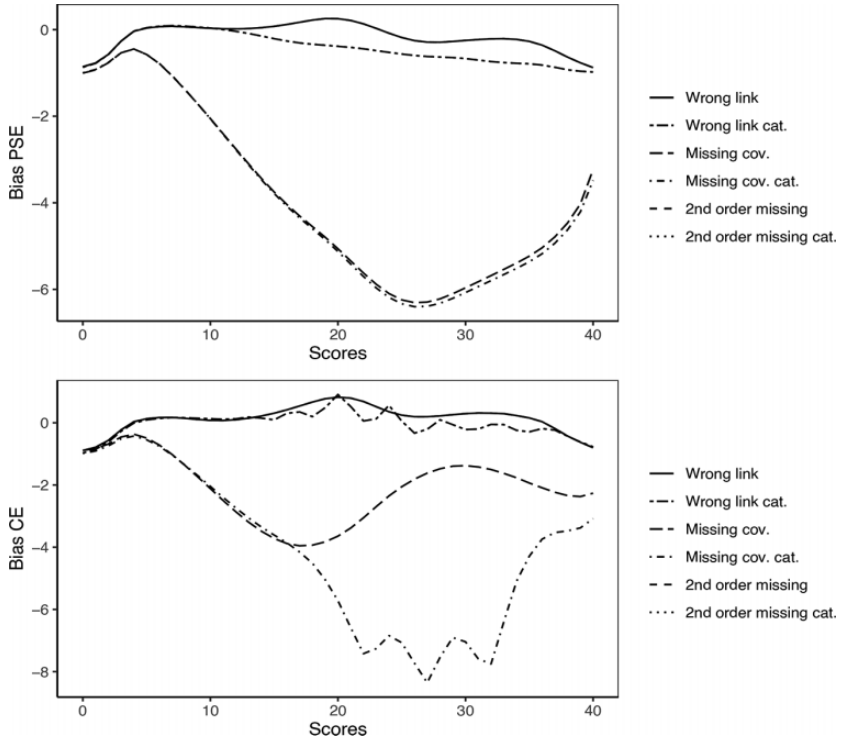


FIGURE 6. The bias of the PS-PSE and PS-CE estimators for $n = 10,000$ test-takers under Simulation Design A, considering both categorized and uncategorized covariates, using a misspecified link function and a missing covariate, respectively, in the propensity score estimation model.

5.4. Simulation Results—Design A

The bias of the PS-PSE and PS-CE estimators is presented in Figure 6. Note that for propensity score models with a misspecified link function and for those that fail to include the second order term, the bias is very similar. Although not illustrated in the figure, their biases practically coincide with the biases of their correctly specified counterparts (the difference is less than 0.01 for each score point). This turns out to be a pattern which is present for both estimators for all considered sample sizes, all evaluation measures, and both simulation designs.

As the upper part of Figure 6 illustrates, the PS-PSE estimators exhibit only a small bias for all scores, with the exception of the KE estimators with a propensity score model that leaves out a covariate. It is also noteworthy that it does not matter whether or not the covariates have been categorized; the biases for all estimators stay similar regardless. The estimators that misspecify the link

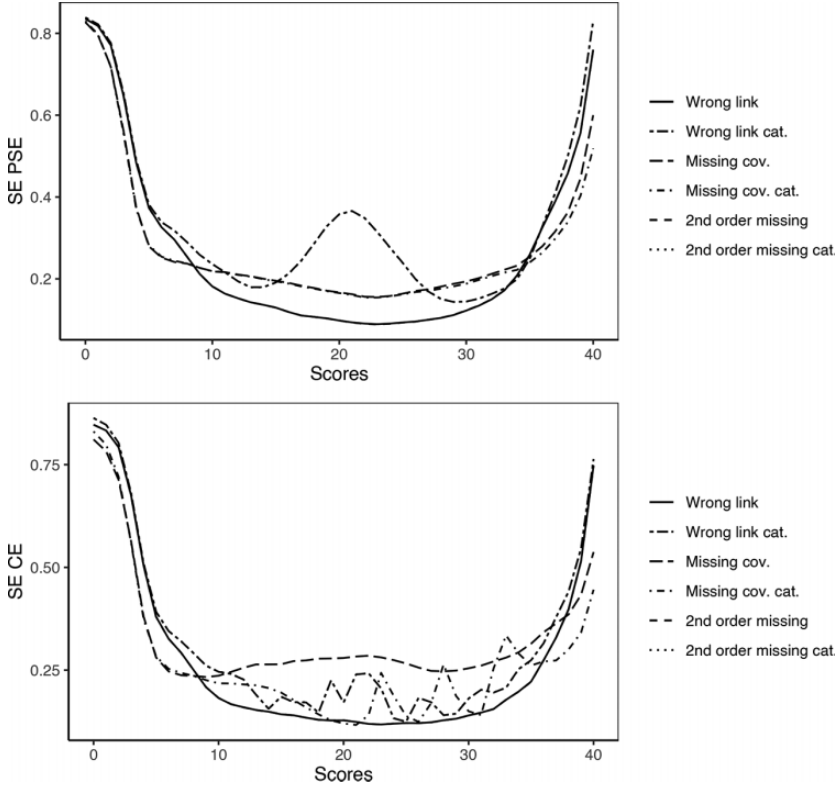


FIGURE 7. The standard error of the PS-PSE and PS-CE estimators for $n = 10,000$ test-takers under Simulation Design A, considering both categorized and uncategorized covariates, using a misspecified link function and a missing covariate, respectively, in the propensity score estimation model.

function and that leaves out the second-order term show the best performance, with differences between them being too small to be discovered in the figure. As these estimators more or less coincide with the estimator using a correctly specified model, the results suggest that the propensity score is successful at balancing the test groups for the PS-PSE estimator.

The lower part of Figure 6 depicts the bias for the PS-CE estimators. For the PS-PSE estimators, misspecifying the link function (and leaving out the second order term) yields small biases across the score range. There is a negligible difference between using categorized and uncategorized covariates in the propensity score model, and the bias increases substantially when a covariate is left out and grows particularly large for categorized covariates.

The *SE* of the PS-PSE and PS-CE estimators is illustrated in Figure 7. Generally, the *SE* is larger in the lower and upper end of the score range

regardless of the type of misspecification. This is due to the sparse data at the most extreme scores. The estimators perform similarly with few exceptions. However, the PS-PSE estimator with a propensity score model that leaves out the second-order term of the categorized covariates yields a slightly larger *SE*, especially in the middle segment of the score scale. For the PS-CE estimators, it instead is the misspecification consisting of a left out covariate that results in such pattern. We remind that the solid curves also represent the results of the estimators with the second-order terms missing, down to a very small difference. The dot dashed curves in the same way represent two types of misspecifications for categorized covariates.

From Design A, we conclude that misspecifying the link function or missing to include a second-order term, for both the original covariates and the categorized versions of them, introduces far less error compared to missing to include a covariate in the propensity score model.

5.5. Simulation Results—Design B

The results of Design B are presented for $n = 10,000$ as the results for $n = 1,000$ and $n = 5,000$ are more or less the same, both in magnitudes of the evaluation measures and in relative performance of the estimators.

The bias of the PS-PSE and PS-CE estimators is displayed in Figure 8. The similarity with the biases in Design A is apparent. Once again, failing to include an important covariate leads to severe bias for both estimators. Especially in the case of the PS-CE estimator with categorized covariates, the results are particularly inaccurate. The estimators with a misspecified link function and those who fail to include the second-order term show robust results in the presence of model misspecification.

The *SE* of the estimators is shown in Figure 9. Both estimators perform similarly for all misspecifications, but with an overall best performance shown in the case of misspecified link functions and with a second order missing, respectively. In contrast to the other estimators, the *SE* of these estimators also drops for the top scores. This could be a meaningful difference since the most critical decisions in many tests, for example, selection tests, are made at the top scores. It should however be noted that the *SEs* are large, especially in the tails.

Similar to Design A, we conclude from Design B that the estimators with an incorrect link function and those that do not include the second-order term are relatively robust. The PS-CE estimator that fails to include one of the categorized covariates shows the overall worst performance. We also observe that the results of Design B are approximately proportional to those of Design A, possibly due to both designs having the same type of covariates (uniformly distributed on the interval $[1, 5]$). However, Design B has a more intricate relationship between the covariates and the propensity score, as well as the covariates and the test scores. As a result, the biases displayed in Design B's results are roughly twice as large

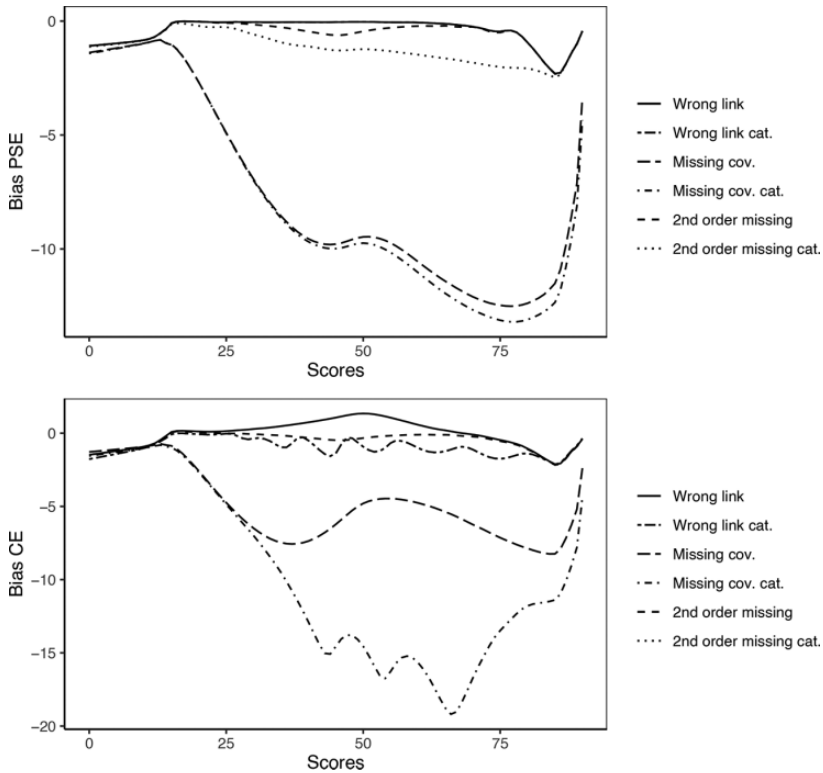


FIGURE 8. The bias of the PS-PSE and PS-CE estimators for $n = 10,000$ test-takers under Simulation Design B, considering both categorized and uncategorized covariates, using a misspecified link function and a missing covariate, respectively, in the propensity score estimation model.

as those seen in Design A, and the *SEs* have also increased. Therefore, the additional complexity in the DGP has amplified the equating error.

6. Discussion

The goal of this study was to investigate how sensitive the equated scores are to model misspecification of the propensity score, when the propensity score is used to equate nonequivalent test groups. It has already been shown in Wallin and Wiberg (2019) that equating with propensity scores has the possibility to reach similar precision and accuracy as equating with an anchor, and superior results compared to equating under a false assumption of equivalent groups. But since the results of Wallin and Wiberg (2019) are based on the assumption that the propensity score is known, which it typically is not in practical research

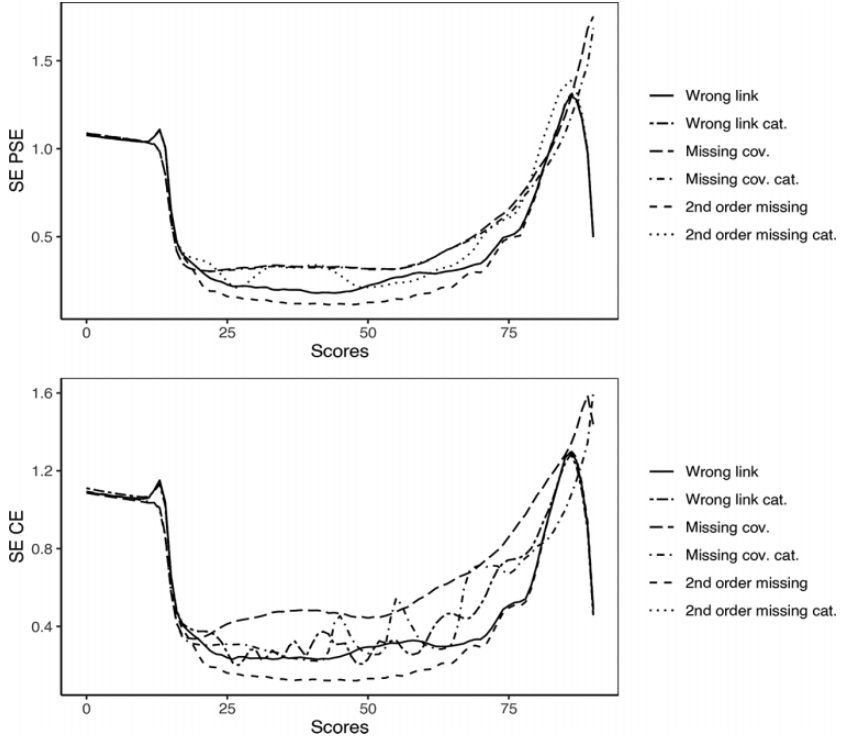


FIGURE 9. The standard error of the PS-PSE and PS-CE estimators for $n = 10,000$ test-takers under Simulation Design B, considering both categorized and uncategorized covariates, using a misspecified link function and a missing covariate, respectively, in the propensity score estimation model.

scenarios, it was crucial to study how sensitive these results are to model misspecification. The propensity score is a useful tool in research as it possesses the desirable feature of being a balancing score, which has led to its widespread application across various domains. However, its high degree of flexibility means that there are numerous modeling options available, emphasizing the need for careful scrutiny to determine when the propensity score can effectively balance test-taker groups and when it falls short.

The propensity score methods explored in this study demonstrate potential as the equated scores remain insensitive to both link function misspecification and the omission of a second-order term in the estimation model. This applies to both linear (Simulation Design A) and nonlinear (Simulation Design B) relationships between covariates and outcomes. Notably, the model misspecifications resulted in a similar bias and SE (in rounded score terms) to the correctly specified models, signifying robustness of the equated scores to such errors in the

propensity score model. On the other hand, the equated scores were negatively affected by a propensity score model that omitted a true confounding covariate. These conclusions remained the same for all considered sample sizes and for both simulation designs. The results therefore clearly point to the importance of using all pertinent information related to latent ability when using the propensity score as a proxy variable. This aligns with earlier research on the propensity score, which indicates that omitting a higher order term that exists in the actual model while estimating the propensity score does not result in biased estimates (Dehejia & Wahba, 1999; Drake, 1993; Stuart, 2010; Waernbaum, 2010, 2012). Incorporating all true confounding variables is linked to the unconfoundedness assumption that forms the foundation of the propensity score method for covariate balancing. Consistent with earlier research, it was found that this aspect is crucial in the equating context as well. As in Waernbaum (2010, 2012), we note that as long as the true propensity score is a function of the misspecified model, unbiased estimation of the parameter of interest is possible. We note that for Design B, the standard errors are fairly large but should be seen in relation to previous research that has showed that equating error and variability is even greater when falsely assuming equivalent groups (Wallin & Wiberg, 2019). A misspecification of the propensity score model when the relationship between the test scores and the covariates is nonlinear is thus a delicate scenario. Since reported scores often are used for individual-level decision making, the current results suggest that future research should carefully study nonlinear cases.

We emphasize that the quality of the ability balancing suggested in this article depends strictly on the quality of the auxiliary information. The restrictions that come with the data at hand need to be evaluated with the identifying Assumptions 1 and 2 in mind. Two examples of restrictions in the empirical data analyzed in this study are the limited amount of covariates and the fact that the variable Age is only available in a categorized version. Since the proposed method has been shown to perform similar to anchor test-based equating for this particular data set (Wallin & Wiberg, 2019), there is reason to believe that the current covariate restrictions have not reversed the results. In the case of propensity score-based equating, we advise seeking input from experts in the subject matter concerning the testing program and test groups that need to be equated. Additionally, we suggest conducting a comprehensive analysis of the associations between the collected covariates and test scores. Since both the propensity score and anchor test score are employed as proxies for ability, they can be evaluated using similar methods.

Some limitations with the current study include the following. We only considered two types of covariates and future studies could consider to expand that. Both by using a propensity score model that is a function of both discrete and continuous covariates and with different dependence structure between them. We however emphasize that the aim of this article was to study propensity score model misspecification, and the misspecifications were thus the main focus and

not different types of covariates. We therefore chose to vary the relationship between the treatment variable, the test scores, and the covariates but not the covariates themselves. On this note, it should be pointed out that Assumptions 1 and 2 are strong, but of similar magnitude to the assumptions underlying NEAT equating. The results in both the original paper by Wallin and Wiberg (2019) and the current article furthermore suggest that there are several realistic test scenarios, where propensity score stratification is a viable technique for a sufficient ability imbalance reduction. It would therefore be of importance to further investigate how sensitive the equating function parameter is to violations of the propensity score assumption. Studying the omission of a true confounder in the propensity score model could be considered a first step toward such analysis, since this violated the unconfoundedness assumption in Assumption 1. A diagnostic tool would in the future be of great use for such analysis. In Online Appendix C, further simulation results are presented, considering both missing data and another case of model assumption violation. These results suggest that the PS-PSE particularly is robust against certain missingness, but that bias is introduced when a subset of test-takers have a true propensity score equal to 1 (or equivalently, equal to 0). These scenarios could, for example, happen when there is an age restriction to the test in question, and certain test-takers were not allowed to take the test in the previous administration. An empirical check of the propensity scores should therefore always be conducted.

It is worth mentioning that the outcomes of Simulation Design B demonstrate a proportional relationship with those of Simulation Design A. This is attributed to the intricate association among the covariates, the treatment variable, and the outcome in Design B, which is more complicated than that in Design A. In addition to these factors, there are testing programs that have access to both covariates and an anchor test. It would therefore be worth investigating if there is any additional gain by using both sources of information to control for ability differences. Incorporating both covariates and anchor test scores has been studied within the NEC design (Albano & Wiberg, 2019; Wiberg & Bränberg, 2015), but never when considering propensity scores. We expect this to improve the results, as demonstrated in the small example in Online Appendix C. Generalizing these results and quantifying the improvement would be a significant contribution to equating nonequivalent groups. Finally, this study has only considered parametric regression models to estimate the propensity score, and other existing methods should be examined in future research.

As a final note, we point out the recent critique that has been raised toward NEAT-based equating in San Martín and González (2022). With data being partially missing by design in the nonequivalent groups designs, the test score distributions, and thus the equating estimator, are not identified. Most methods, including the methods studied in this article, make identifying assumptions to estimate the score distributions. An alternative approach, suggested in San Martín and González (2022), is to use the theory of partial identification (Manski, 2009) to

define identification regions for the equating function. This is a new perspective that we believe sheds light to the discussion on whether or not equating has any potential to report fair scores under nonequivalent groups designs, see, for example, Bolsinova and Maris (2016). Their approach could also serve as a useful tool to investigate the sensitivity of the identifying assumptions presented in this article.


Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This work was supported by Vetenskapsrådet (2020-06484) and Marianne och Marcus Wallenbergs Stiftelse (2019.0129).

ORCID iD

Gabriel Wallin  <https://orcid.org/0000-0002-7930-6701>

References

- Akaike, M. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Albano, A. D., & Wiberg, M. (2019). Linking with external covariates: Examining accuracy by anchor type, test length, ability difference, and sample size. *Applied Psychological Measurement*, 43(8), 597–610.
- Altıntaş, 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 with the package kequate. *Journal of Statistical Software*, 55(6), 1–25.
- Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement*, 11(3), 291–300.
- Augurzyk, B., & Schmidt, C. (2001). Iza discussion paper 271, Institute for the Study of Labor (IZA).
- 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.
- Bolsinova, M., & Maris, G. (2016). Can IRT solve the missing data problem in test equating? *Frontiers in Psychology*, 6, 1956.
- 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.

- Braun, H., & Holland, P. (1982). Observed-score test equating: A mathematical analysis of some ets equating procedures. In P. Holland & D. Rubin (eds.), *Test equating*, Vol. 1 (pp. 9–49). Academic Press.
- Brennan, R. L., & Kolen, M. J. (1987a). Reply to angoff. *Applied Psychological Measurement*, 11, 301–306.
- Brennan, R. L., & Kolen, M. J. (1987b). Some practical issues in equating. *Applied Psychological Measurement* 11(3): 279–290.
- Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1990). Equating achievement tests using samples matched on ability. *ETS Research Report Series*, 1990(1), i–58.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1), 1–15.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448), 1053–1062.
- Dorans, N., & 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). Research & Development.
- Dorans, N., & Holland, P. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.
- Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4), 1231–1236.
- González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.
- Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, 40(3), 254–273.
- Hernan, M., & Robins, J. (2020). *Causal inference*. Chapman & Hall/CRC.
- Holland, P., & Thayer, D. (1989). *The kernel method of equating score distributions*. Technical report, Educational Testing Service.
- Hsu, T.-c., Wu, K.-l., Yu, J.-y. W., & Lee, M.-y. (2002). Exploring the feasibility of collateral information test equating. *International Journal of Testing*, 2(1), 1–14.
- Invalsi. (2013). *Rilevazioni nazionali sugli apprendimenti 2012-13*. Technical report, Invalsi Publishing.
- Kolen, M. J. (1990). Does matching in equating work: A discussion. *Applied Measurement in Education*, 3(1), 97–104.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
- Lee, Y., & von Davier, A. (2011). Equating through alternative kernels. In A. von Davier (ed.), *Statistical models for test equating, scaling, and linking*, Vol. 1 (pp. 159–173). Springer.
- Liou, M., Cheng, P. E., & Li, M.-Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, 25(2), 197–207.
- Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1), 73–95.
- Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, 40(3), 227–253.
- Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

- Moses, T., Deng, W., & Zhang, Y.-L. (2010). The use of two anchors in Nonequivalent Groups With Anchor Test (NEAT) equating. *ETS Research Report Series*, 2010(2), i–33.
- 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.
- Powers, S. J. (2010). *Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions*. The University of Iowa.
- 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.
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 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(387), 516–524.
- San Martín, E., & González, J. (2022). A critical view on the neat equating design: Statistical modeling and identifiability problems. *Journal of Educational and Behavioral Statistics*, 47(4), 10769986221090609.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Sinharay, S., & Holland, P. (2010a). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, 47(3), 261–285.
- Sinharay, S., & Holland, P. W. (2010b). The missing data assumptions of the neat design and their implications for test equating. *Psychometrika*, 75(2), 309–327.
- 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.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1.
- von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, 78(4), 605–623.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, 41(1), 15–32.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. Springer.
- Waernbaum, I. (2010). Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference*, 140(7), 1948–1956.
- Waernbaum, I. (2012). Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Statistics in Medicine*, 31(15), 1572–1581.

- Wallin, G., Häggström, J., & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating? *Applied Psychological Measurement*, 45, 518–535.
- Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, 44(4), 390–414.
- 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.
- Wright, N. K., & Dorans, N. J. (1993). Using the selection variable for matching or equating 1, 2. *ETS Research Report Series*, 1993(1), i–22.

Authors

GABRIEL WALLIN is a visiting fellow at the Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; e-mail: g.a.wallin@lse.ac.uk, and a postdoctoral researcher at the Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, SE-901 87 Umeå, Sweden. His research interests are centered on statistical models, machine learning algorithms, and other quantitative methods that aid fairness and interpretability in the social and behavioral sciences. Topics include model-based clustering, multivariate outlier detection, and exploratory factor analysis.

MARIE WIBERG is a professor at the Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, SE-901 87 Umeå, Sweden; e-mail: marie.wiberg@umu.se. Her research interests include educational measurement and psychometrics in general, especially test equating, parametric and nonparametric item response theory, and international large-scale assessments.

Manuscript received December 9, 2021

Revision received November 30, 2022

Accepted January 29, 2023