Article

Journal of Educational and Behavioral Statistics
2019, Vol. 44, No. 4, pp. 390–414
DOI: 10.3102/1076998619838226
Article reuse guidelines: sagepub.com/journals-permissions

Kernel Equating Using Propensity Scores for Nonequivalent Groups

Gabriel Wallin
Marie Wiberg
Umeå University

When equating two test forms, the equated scores will be biased if the test groups differ in ability. To adjust for the ability imbalance between nonequivalent groups, a set of common items is often used. When no common items are available, it has been suggested to use covariates correlated with the test scores instead. In this article, we reduce the covariates to a propensity score and equate the test forms with respect to this score. The propensity score is incorporated within the kernel equating framework using poststratification and chained equating. The methods are evaluated using real college admissions test data and through a simulation study. The results show that propensity scores give an increased equating precision in comparison with the equivalent groups design and a smaller mean squared error than by using the covariates directly. Practical implications are also discussed.

Keywords: *kernel equating; background variables; nonequivalent groups; NEC design; propensity scores*

1. Introduction

Scores from different versions of a test can be compared only after equating (González & Wiberg, 2017). Having items common to both versions and test takers who took both versions are important factors in the choice of methods and designs for equating. For example, the equivalent groups (EG) design requires test groups that have identical distributions of ability. In the design with nonequivalent groups with an anchor (NEAT), score adjustment is based on the items common to the two versions. Since test groups in large-scale assessments could not always be considered equivalent, the NEAT design is preferable in many situations. Problems arise when the groups are not equivalent and there are no common items, for example, the INVALSI test (INVALSI, 2013) and the Armed Services Vocational Aptitude Battery (Quenette, Nicewander, & Thomasson, 2006). Background variables may be used as substitutes for the score on the common items through a nonequivalent groups with covariates (NEC) design

(Wiberg & Bränberg, 2015). In the NEC design there is no anchor test available, so covariates that are correlated with the test scores are used instead. In Wiberg and Bränberg (2015), the test takers are categorized through a number of covariates that correlate highly with the test scores as a way to replace anchor test scores in a novel design built on poststratification equating (PSE). However, they did not consider the possibility of using propensity scores (Rosenbaum & Rubin, 1983, 1984) nor did they examine the possibility of using a chained equating (CE) approach within the NEC design. Their method also requires continuous variables to be converted to categorical variables. A problem with their approach is that the number of covariate categories quickly expands with an increasing number of covariates, making the number of categories that have few or even no observations, from one or both tests, to proliferate. We handle this problem by using propensity scores instead of the covariates.

The general use of collateral information has been suggested several times in equating research. Kolen (1990); Cook, Eignor, and Schmitt (1990); and Wright and Dorans (1993) used covariates in matched sampling to adjust for differences between the test groups. Liou, Cheng, and Li (2001) replaced the anchor test with covariates, Bränberg and Wiberg (2011) incorporated covariates in linear equating, and Hsu, Wu, Yu, and Lee (2009) used covariates in item response theory (IRT) true-score equating. This article differs from these papers by applying propensity scores as a way to use collateral information to control for nonequivalent groups.

To use propensity scores in test equating is not a new idea, and it was first proposed by Livingston, Dorans, and Wright (1990). Since then, there have been several suggestions on how to use it within equating. Yu, Livingston, Larkin, and Bonett (2004) and Paek, Liu, and Oh (2006) used propensity scores to match samples, and Sungworn (2009) used it to improve the traditional PSE. Later, Moses, Deng, and Zhang (2010) used propensity scores to combine two anchor test scores for use in PSE, and Powers (2010) used propensity scores within CE, frequency estimation, IRT true-score equating, and IRT observed-score equating. Recently, Longford (2015) proposed an equating method that made use of matching with either inverse proportional weighting or matched pairs based on propensity scores, and Haberman (2015) used propensity scores to create pseudoequivalent groups from nonequivalent groups. However, none of the above mentioned studies used propensity scores within the kernel equating framework as proposed here.

Motivated by the fact that kernel equating has become a well-established framework for both theorists and practitioners, we apply a kernel equating estimator that uses propensity scores within the NEC design and demonstrate both a PSE approach and a CE approach with real and simulated test data. Several studies have indicated that CE is less biased than PSE when there are large differences in ability between the test groups (e.g., Livingston, Dorans, & Wright, 1990; Wang, Lee, Brennan, & Kolen, 2006), but von Davier, Holland,

and Thayer (2004a) showed that the two approaches under certain conditions are equivalent. In this article, we compare CE and PSE under the NEC design. Previous studies have also shown that the choice of equating transformation is less important when the anchor test shows close resemblance to the overall test. This result is of particular interest to investigate when covariates replace the anchor, since covariates almost always are less correlated with the test scores than the anchor test scores. The proposed approaches are empirically examined using data from a college admission test and are compared to kernel equating within the EG and the NEAT design. The choice of collateral information is important, and the objective is to incorporate variables that both correlate highly with the test scores and that can explain differences between the groups of test takers. Other studies of the test used in the empirical illustration of this article have shown that gender, education, and age correlate with the test scores (Bränberg, Henriksson, Nyquist, & Wedman, 1990), and in this article, both gender and age were used together with the scores from a different part of the test. It should though be noted that the logistics in collecting variables are commonly limited.

The structure of this article is as follows: First, the NEC design is described in general, followed by an introduction to propensity scores. Next, the kernel equating framework is presented, and two approaches for kernel equating using propensity scores are proposed. The proposed methods are then illustrated and examined in an empirical illustration and a simulation study, which is followed by some concluding remarks.

2. NEC Design

Equating nonequivalent test groups means having to adjust for two sources of bias: differences in difficulty of the forms and differences in ability of the test groups. A suitable equating transformation handles the former, but since the latter is unobserved, some proxy for ability is needed. The most common proxy is an anchor test, for which there exist several NEAT design equating transformations. However, not all testing programs can include an anchor. Instead, there might be other background information of the test takers available. This is the scenario of the NEC design, where the underlying assumption is that if the test groups are equivalent conditionally on background information, they will be only randomly different from each other in terms of ability.

To formalize the NEC design, assume that we have a sample of size n_P from a population P and a sample of size n_Q from another population Q. The sample from P has been administered test form X, and the sample from Q has been administered test form Y, where X and Y represent the same assessment test but contain different sets of items. For each test taker, an observation of either a score X if administered test form X or a score Y if administered test form Y is registered.

There are no common items available; however, a set of background variables $\mathbf{D} = (D_1, \ldots, D_m)$ is observed for each test taker. The background variables are referred to as covariates and have the purpose of controlling for differences in ability between the test groups (Wiberg & Bränberg, 2015). The aim is to control for all the confounding covariates, that is, those that affect both the treatment assignment and the test scores. This will balance the covariate distribution in the test groups, making them only randomly different from each other. In that sense, covariates are used as a proxy for ability, just as an anchor might be.

2.1. Propensity Scores

The covariates within the NEC design will be used in a univariate composite called the propensity score. For each test taker, the propensity score $e(\mathbf{D})$ specifies the conditional probability of being assigned to a particular treatment (i.e., test form) given the covariate vector $\mathbf{D}$ (Rosenbaum & Rubin, 1983). Letting Z denote a treatment variable equal to 1 if test form Y (the active treatment) is administered and 0 if test form X (the control treatment) is administered, the propensity score is defined as $e(\mathbf{D}) = \Pr(Z = 1|\,\mathbf{D}) = E(Z|\,\mathbf{D})$. We will use the fact that the propensity score is a balancing score, meaning that if $\mathbf{D}$ contains every confounder of the (X, Y) and Z relationship, it is sufficient to control for $e(\mathbf{D})$ to create covariate balance in the test groups. This will be formalized in the next section in two slightly different ways, depending on whether it is PSE or CE that is applied.

The propensity score for each test taker is unknown since the administration of test forms is not randomized. It can be estimated in several ways, most commonly using logistic regression. A practical guideline is to include every covariate in the estimation model that is strongly correlated with the test scores. Once the estimated propensity score is obtained, Rosenbaum and Rubin (1984) proposed subdividing the vector of propensity scores into strata based on the percentiles. This approach suggests that test takers with propensity scores falling into the same stratum are equivalent with respect to ability. Thus, the number of strata should be chosen such that the test groups are homogeneous within each stratum in terms of covariate distribution.

3. Kernel Equating

Kernel equating aims to equate X to Y in the target population T. For the nonequivalent groups designs, T is defined as a mixture of populations P and Q, that is, $T = wP + (1 - w)Q$, where $0 \leq w \leq 1$. The equating procedure comprises the following five steps: (1) presmoothing, (2) estimation of the score probabilities, (3) continuization, (4) equating, and (5) calculation of evaluation measures (von Davier, Holland, & Thayer, 2004b, pp. 45–47; see also González & Wiberg, 2017). Let the realizations of X and Y be denoted $x_j, j = 1, \ldots, J$, 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 respective probabilities of a randomly selected test taker in T scoring x_j on test form X and y_k on test form Y. The cumulative distribution functions (CDFs) of X and Y in T are denoted $F(x) = \Pr(X \leq x|T)$ and $G(y) = \Pr(Y \leq y|T)$, respectively.

Kernel equating uses the equipercentile transformation to equate X to Y, which states that an equivalent score y on form Y to a score x on form X is given by

$$y = \varphi_Y(x) = G^{-1}(F(x)). \tag{1}$$

Since $F(\cdot)$ and $G(\cdot)$ need to be continuous for $\varphi(\cdot)$ to be properly defined, kernel equating implements a, usually Gaussian, kernel to continuize the score CDFs. With this purpose, let $\mathbf{r} = (r_1, \ldots, r_J)^t$, $\Phi(\cdot)$ denote the standard normal distribution function, $\mu_X = \sum_j x_j r_j$ denote the mean of X in population T, $a_X = \sqrt{\sigma_X^2/(\sigma_X^2 + h_X^2)}$, σ_X^2 denote the variance of X in T, and $h_X > 0$ denote a smoothing parameter. Using the Gaussian kernel, the continuized X score CDF is defined as

$$F_{h_X}(x; \mathbf{r}) = \Pr\Big(X(h_X) \leq x\Big) = \sum_j r_j \Phi\left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X}\right). \tag{2}$$

The continuization of the Y score distribution to obtain $G_{h_Y}(y; \mathbf{s})$, $\mathbf{s} = (s_1, \ldots, s_K)^t$, is done similarly. The so-called bandwidths h_X and h_Y can be selected in several ways, for example, by double smoothing (Häggström & Wiberg, 2014) or by minimizing a penalty function (von Davier, 2013; von Davier et al., 2004b, p. 63). Thus, it is only the score probabilities that need to be estimated for it to be possible to calculate the continuized CDFs, and thereby form the equipercentile transformation given in Equation 1. For this purpose, the estimated propensity scores will be used, as explained in the following section.

In the final step of kernel equating, the estimated equating transformation is evaluated. The most common evaluation measure is the standard error of equating (SEE), which in kernel equating is given by

$$\mathrm{SEE}_Y(x) = ||\mathbf{J}_{\varphi_Y} \mathbf{J}_{\mathrm{DF}} \mathbf{C}||, \tag{3}$$

where $||\cdot||$ denotes the Euclidean norm, $\mathbf{J}_{\varphi_Y}$ is the Jacobian of the equating transformation, $\mathbf{J}_{\mathrm{DF}}$ is the Jacobian of the design function mapping the estimated score distributions into estimates of $\mathbf{r}$ and $\mathbf{s}$, and $\mathbf{C}$ is a matrix related to the covariance of the estimated score distributions. The explicit expressions of the two Jacobian matrices for the estimators presented in this article are very similar to those when using anchor scores, see Appendix B for the details.

4. Kernel Equating Using Propensity Scores

4.1. A PSE Approach

The first equating transformation of our proposal takes a PSE approach and is abbreviated PSE NEC PS. In this approach, we define the distributions of X and Y in T as

$$r_j = \Pr(X = x_j|T) = wr_{Pj} + (1 - w)r_{Qj}, \tag{4}$$

and

$$s_k = \Pr(Y = y_k|T) = ws_{Pk} + (1 - w)s_{Qk}, \tag{5}$$

where $r_{Pj} = \Pr(X = x_j|P)$, $r_{Qj} = \Pr(X = x_j|Q)$, $s_{Pk} = \Pr(Y = y_k|P)$, and $s_{Qk} = \Pr(Y = y_k|Q)$ are the score distributions of X and Y in populations P and Q, respectively. In order to define estimators of r_j and s_k, denote the stratified propensity score for strata l, $l = 1, \ldots, L$, by $e_{Xl}(\mathbf{D})$ for population P and $e_{Yl}(\mathbf{D})$ for population Q. The joint distributions of $\left(X, e_{Xl}(\mathbf{D})\right)$ and $\left(Y, e_{Yl}(\mathbf{D})\right)$ are denoted as $p_{jl} = \Pr(X = x_j, e_{Xl}(\mathbf{D}) = e_{Xl}(\mathbf{d})|P)$ and $q_{kl} = \Pr\left(Y = y_k, e_{Yl}(\mathbf{D}) = e_{Yl}(\mathbf{d})|Q\right)$, respectively, where $\mathbf{d}$ denotes the observed value of $\mathbf{D}$. The probabilities r_{Pj} and s_{Qk} are estimated by

$$\hat{r}_{Pj} = \sum_l \hat{p}_{jl} \text{ and } \hat{s}_{Qk} = \sum_l \hat{q}_{kl}, \tag{6}$$

where $\hat{p}_{jl}$ and $\hat{q}_{kl}$, for example, could be estimates from a presmoothing model such as a log-linear model. By design, there is no data to estimate r_{Qj} and s_{Pk} directly. To derive estimators of these quantities, we present the following PSE NEC PS assumptions about the propensity score:

$$X \perp Z|e(\boldsymbol{D}), \tag{A.1}$$

$$Y \perp Z|e(\boldsymbol{D}), \tag{A.2}$$

$$0 < e(\boldsymbol{D}) < 1, \tag{A.3}$$

where $\perp$ denotes statistical independence. In causal inference, Assumptions A.1 and A.2 are known as the unconfoundedness assumption, and Assumption A.3 is called the overlap assumption (Abadie & Imbens, 2006). Rosenbaum and Rubin (1983) proved that A.1 and A.2 holds if $\mathbf{D}$ contains every confounding covariate. Since they are untestable assumptions, the aim should be to include all measured covariates in the propensity score estimation model that affects both treatment and scores.

Assumptions A.1 and A.2 implies that the conditional distribution of X given $e(\mathbf{D})$ and Y given $e(\mathbf{D})$ is the same in population P and Q, respectively. This makes it possible to estimate the missing quantities in Equations 4 and 5 by

$$\hat{r}_{Qj} = \sum_l \left(\frac{\hat{p}_{jl}}{\sum_j \hat{p}_{jl}} \cdot \sum_k \hat{q}_{kl} \right) \text{ and } \hat{s}_{Pk} = \sum_l \left(\frac{\hat{q}_{kl}}{\sum_k \hat{q}_{kl}} \cdot \sum_j \hat{p}_{jl} \right), \tag{7}$$

thus making every unknown term of r_j and s_k possible to estimate.

Once $\hat{r}_j$ is calculated, it is plugged into Equation 2 to estimate F_{h_X}. The CDF G_{h_Y} is obtained analogously using $\hat{s}_k$. The PSE NEC PS equating transformation that equates X to Y in population T is obtained by composing the equipercentile transformation from $\hat{F}_{h_X}$ and $\hat{G}_{h_Y}$:

$$\hat{\varphi}_{Y(PSE)}(x) = \varphi_{Y(PSE)}(x; \widehat{\mathbf{r}}, \widehat{\mathbf{s}}) = G_{h_Y}^{-1}\left(F_{h_X}(x; \widehat{\mathbf{r}}); \widehat{\mathbf{s}}\right) = \hat{G}_{h_Y}^{-1}\left(\hat{F}_{h_X}(x)\right). \tag{8}$$

4.2. A CE Approach

CE using propensity scores is referred to as CE NEC PS. This approach first links X to $e_{Xl}(\mathbf{D})$ in population P and then $e_{Yl}(\mathbf{D})$ to Y in population Q. To make this procedure valid as an observed-score equating method, we present a set of assumptions underlying CE NEC PS. First, let $H\left(e(\mathbf{d})\right) = \Pr\left(e(\mathbf{D}) \leq e(\mathbf{d})|T\right)$, $H_P\left(e(\mathbf{d})\right) = \Pr\left(e(\mathbf{D}) \leq e(\mathbf{d})|P\right)$, $H_Q\left(e(\mathbf{d})\right) = \Pr\left(e(\mathbf{D}) \leq e(\mathbf{d})|Q\right)$, $F_P(x) = \Pr(X \leq x|P)$, and $G_Q(y) = \Pr(Y \leq y|Q)$. The CE NEC PS assumptions are

$$H_P^{-1}(F_P(x)) = H^{-1}(F(x)), \tag{B.1}$$

$$G_Q^{-1}(H_Q(e(\mathbf{d}))) = G^{-1}(H(e(\mathbf{d}))). \tag{B.2}$$

Assumptions B.1 and B.2 state that the links from X to $e_{Xl}(\mathbf{D})$ and from $e_{Yl}(\mathbf{D})$ to Y are population invariant for any T of the form $T = wP + (1-w)Q$. From Assumptions B.1 and B.2, it follows that

$$F(x) = H\left(H_P^{-1}\left(F_P(x)\right)\right)$$

and

$$G^{-1}(y) = G_Q^{-1}\left(H_Q\left(H^{-1}(y)\right)\right),$$

meaning that equipercentile transformation of Equation 1 can be formed by

$$G^{-1}\left(F(x)\right) = G_Q^{-1}\left(H_Q\left(H^{-1}\left(H\left(H_P^{-1}\left(F_P(x)\right)\right)\right)\right)\right) = G_Q^{-1}\left(H_Q\left(H_P^{-1}\left(F_P(x)\right)\right)\right).$$

The score probabilities needed for this approach are $\mathbf{r}_p = (r_{P1}, \ldots, r_{PJ})^t$, $\mathbf{s}_Q = (s_{Q1}, \ldots, s_{QK})^t$, $\mathbf{t}_P = (t_{P1}, \ldots, t_{PL})^t$, and $\mathbf{t}_Q = (t_{Q1}, \ldots, t_{QL})^t$, where $t_{Pl} = \Pr\left(e_{Xl}(\mathbf{D}) = e_{Xl}(\mathbf{d})|P\right)$ and $t_{Ql} = \Pr\left(e_{Yl}(\mathbf{D}) = e_{Yl}(\mathbf{d})|Q\right)$. The probabilities r_{Pj}

and s_{Qk} are estimated by Equation 6, and the probabilities t_{Pl} and t_{Ql} by $\hat{t}_{Pl} = \sum_j \hat{p}_{jl}$ and $\hat{t}_{Ql} = \sum_k \hat{q}_{kl}$. These four sets of probabilities are used to estimate $F_P(\cdot), G_Q(\cdot), H_P(\cdot),$ and $H_Q(\cdot)$ using analogous versions of Equation 2. This yields the estimated CDFs: $F_{h_P}(x;\widehat{\mathbf{r}}_P) = \hat{F}_{h_P}(x)$, $G_{h_Q}(y;\widehat{\mathbf{s}}_Q) = \hat{G}_{h_Q}(y)$, $H_{h_{e_{Xl}}}\left(e_{Xl}(\mathbf{d});\widehat{\mathbf{t}}_P\right) = \hat{H}_{h_{e_{Xl}}}\left(e_{Xl}(\mathbf{d})\right)$, and $H_{h_{e_{Yl}}}\left(e_{Yl}(\mathbf{d});\widehat{\mathbf{s}}_Q\right) = \hat{H}_{h_{e_{Yl}}}\left(e_{Yl}(\mathbf{d})\right)$. With this notation introduced, it follows that the CE NEC PS estimator is given by

$$\hat{\varphi}_{Y(CE)}(x) = \varphi_{Y(CE)}(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{t}}_Q,\hat{\boldsymbol{s}}_Q) = \hat{G}^{-1}_{h_Q}\left(\hat{H}_{h_{e_{Yl}}}\left(\hat{H}^{-1}_{h_{e_{Xl}}}\left(\hat{F}_{h_P}(x)\right)\right)\right). \tag{9}$$

5. Empirical Illustration

Data from the Swedish Scholastic Assessment Test (SweSAT) for college admissions were used to illustrate the suggested approaches for incorporating propensity scores within the NEC design in kernel equating. The SweSAT is a paper and pencil test with 160 multiple-choice binary-scored items. It consists of a quantitative section of 80 items and a verbal section of 80 items that are equated separately. The SweSAT is given twice a year and has only recently included an anchor test. Previously, equating was based on a set of covariates, for details, see Lyrén and Hambleton (2011). The empirical illustration was carried out in R (R Core Development Team, 2016) with the kernel equating package kequate (Andersson, Bränberg, & Wiberg, 2013).

Two consecutive administrations of the quantitative section were equated, where the new test form X was equated to the old test form Y. The same raw test score data material was used as in Wiberg and Bränberg (2015), and both test forms were taken by 14,644 test takers. The sample was divided in half, and it was made sure that the covariate distributions differed between the test groups. The anchor test had not been implemented yet for the analyzed administrations, so a 24-item anchor test was constructed through the selection of 12 items from both test administrations. This was possible since the original data consisted of test takers who had taken both forms. The within-form scores have means 39.34 and 43.32 and standard deviations 11.80 and 12.65. The empirical score distributions are illustrated in Appendix A.

The covariates used in the analysis were test scores from the verbal section (with range 0–80), age, and gender, as explained in the introduction. When equating the SweSAT using the whole covariate vector, that is, when $e(\mathbf{D}) = \mathbf{D}$ (referred to as raw covariates), the test scores from the verbal section of SweSAT (Verb) were grouped into four strata: [0–30], [31–40], [41–50], and [51–80]. This is in line with the grouping used in Wiberg and Bränberg (2015). The variable Age was reported only after been categorized into the following

TABLE 1.
Summary of the Covariates Verb (Original Version), Age, and Gender Together With the Anchor Test Used in the Empirical Illustration

	Verb	Age	Gender	Anchor
Correlation to Y	0.48	−0.14	0.26	0.81
Correlation to X	0.52	−0.13	0.28	0.81
Mean	43.91 (39.35)	1 (1)	0.42 (0.53)	12.17 (10.55)
Standard deviation	12.08 (11.56)	2 (2)	0.49 (0.50)	4.59 (4.64)

Note. 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.

strata: [0–20], [21–24], [25–29], [30–39], and 40 or older. The covariates Verb, Age, and Gender together with the anchor scores are summarized in Table 1.

5.1. Equating SweSAT using Propensity Scores

PSE NEC PS and CE NEC PS were compared to PSE using the raw covariates (referred to as PSE RAW COV), PSE and CE within a NEAT design (referred to as PSE NEAT and CE NEAT, respectively), and to equating within the EG design. Firstly, logistic regression was used to estimate the propensity scores by predicting group membership for the test takers. The propensity score estimation model included all covariates (with Verb not categorized) without higher order terms or interactions. Figure 1 displays the histograms of the estimated propensity scores in the two forms.

The estimated propensity scores from the fitted model were divided into strata based on the percentiles. For our purpose, the estimation model was not assessed in terms of goodness of fit but rather by checking the covariate balance in the strata. A common balance measure is the absolute standardized mean difference (ASMD), where a difference less than 0.1 is often regarded as a sign of balance (Austin, 2008). The number of strata was set so that this was achieved for every covariate for as large fraction of strata as possible. In our case, this resulted in 10 strata. The ranges of the ASMD for the covariates Verb, Age, and Gender were [0.003, 0.386], [0.007, 0.258], and [0.005, 0.369], and the number of strata with an ASMD below 0.1 were 5 of 10, 7 of 10, and 4 of 10, respectively. By supplementing the propensity model with suitable interactions, the number of large ASMDs could be reduced. See Appendix D for a table of all ASMDs. The equating results were subject to a sensitivity analysis to make sure that the equated scores were insensitive to a small change of the number of strata.

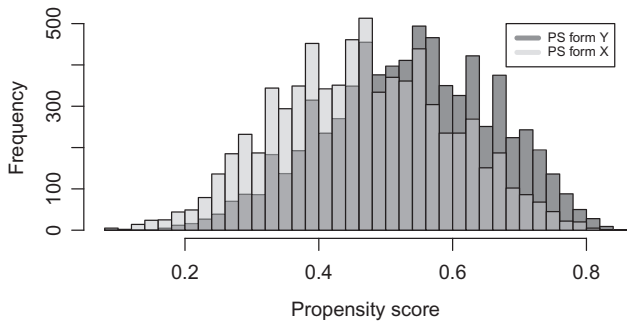

FIGURE 1. *The distribution of the propensity scores for test form* X *(PS form* X*) and* Y *(PS form* Y*). PS = propensity scores.*

Bivariate smoothing of the $\left(X, e_{Xl}(\mathbf{D})\right)$ and $\left(Y, e_{Yl}(\mathbf{D})\right)$ distributions was conducted using log-linear models. For each data design considered, the best model was chosen by evaluating the Bayesian information criterion (Schwarz, 1978). Candidate models with small alterations to the final models were also considered but resulted in only small differences in terms of equated scores and SEE. The included moments together with the chosen bandwidths are given in Appendix C, described for each examined data collection design.

Using Equations 6 and 7, the marginal probabilities r_{Pj}, r_{Qj}, s_{Qk}, and s_{Pk} were estimated. The weight w in Equations 4 and 5 was set to 0.5. The estimated discrete score distributions of X and Y in the target distribution were continuized using a Gaussian kernel and so were the distributions of $e_{Xl}(\mathbf{D})$ and $e_{Yl}(\mathbf{D})$. The bandwidths were selected by minimizing the penalty function given in von Davier, Holland, and Thayer (2004b, p. 63).

5.2. Results of Empirical Illustration

In Figure 2, the difference between the equated scores and the raw scores is plotted for every considered data collection design. All of the PSE approaches are relatively close to each other in terms of equated scores, CE NEC PS deviates from the overall pattern, and CE NEAT is systematically lower than PSE NEAT. The EG design results in very different equated scores compared to those of the PSE methods. The overall largest difference is between equating under the EG design and equating using the CE NEC PS approach, with a maximum score difference of 7.55. The equated scores thus seem more affected by the choice of equating transformation than the choice of instrument to balance the groups.

Figure 3 compares the SEE between the different designs. For most part of the score scale, the differences in SEE are small. For the NEAT design

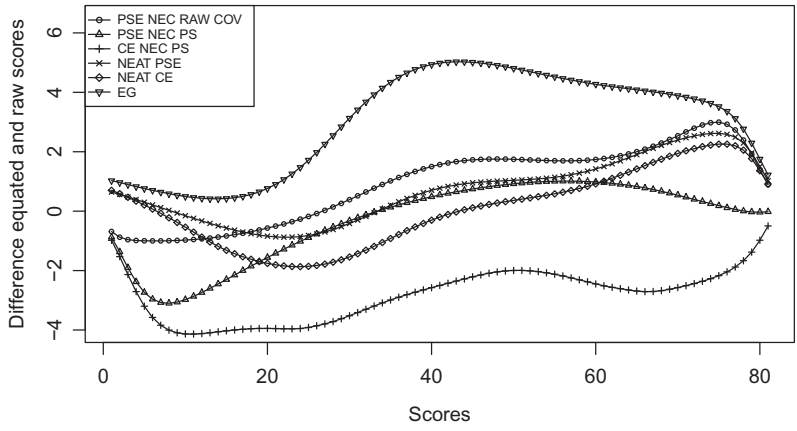

FIGURE 2. *The difference between the equated and raw score, using PSE NEC RAW COV, PSE NEC PS, CE NEC PS, NEAT PSE, NEAT CE, and the EG design. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating; NEAT = nonequivalent groups with an anchor; EG = equivalent groups.*

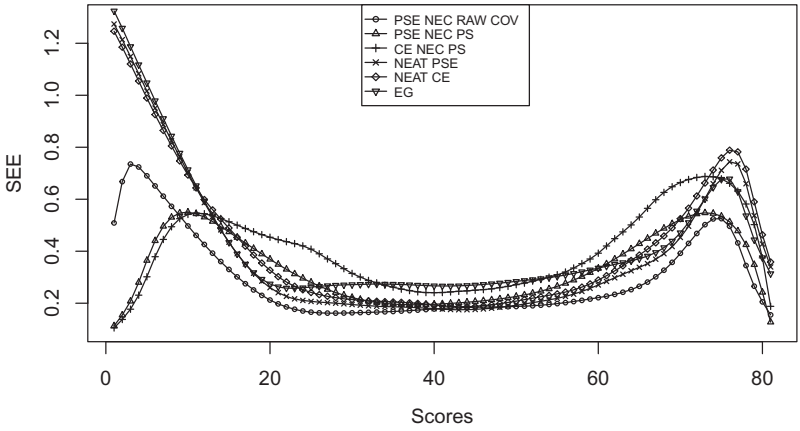

FIGURE 3. *The SEE using PSE NEC RAW COV, PSE NEC PS, CE NEC PS, NEAT PSE, NEAT CE, and the EG design. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating; NEAT = nonequivalent groups with an anchor; EG = equivalent groups; SEE = standard error of equating.*

transformations, the performance is very similar. For the NEC designs, using the raw covariates resulted in the smallest SEE for most of the scores, and CE NEC PS performed somewhat worse in comparison with the other two NEC transformations. It is also evident that the NEC design transformations showed a greater

stability in the tails in comparison with the EG and NEAT design transformations, with smaller SEE values for the lowest and highest scores. EG and CE NEC PS are most dissimilar, with a maximum SEE difference of 1.22 seen in the lower tail of the score scale.

6. Simulation Study

The finite sample properties of PSE NEC PS, CE NEC PS, and PSE NEC RAW COV were examined through a simulation study. Background covariates were used to define the treatment assignment and test scores on two forms X and Y. In this way, the covariates acted as true confounders which made it possible to evaluate the effectiveness of the suggested equating estimators in reducing the bias induced by the covariates. Test scores were generated using the idea of potential outcomes from causal analysis. This means that every test taker had a score on each form: the potential outcome if taken form X and the potential outcome if taken form Y. Both outcomes were generated for every test taker, and the treatment assignment was used to define the actual observed score.

To make the comparisons as fair as possible, no presmoothing was conducted. A Gaussian kernel was employed, and the bandwidth for the X score distribution was the result of minimizing $\sum_j \left(\hat{r}_j - \hat{F}'_{h_X}(x_j)\right)^2$, where $\hat{F}'_{h_X}$ denotes the derivative of $\hat{F}_{h_X}$. The bandwidth for the Y score distribution was selected analogously. A reference equating transformation was defined for each estimator, symbolizing the true relationship between X and Y. These references were formed in the same way as the estimators but used the potential scores on X and Y instead. They are considered to represent the true equating transformations for each respective setting since they involve no missing data. In this way, the missing-data mechanisms of real test data are reflected in the simulation study. This is also in line with the proposed evaluation technique for the NEAT design suggested by Sinharay and Holland (2010).

The simulation study considered two setups to involve different sets of covariates. In the first setup, two continuous covariates were generated, and in the second setup, two discrete covariates were generated. The R package *kequate* (Andersson et al., 2013) was used to equate the test forms.

6.1. Simulation Design—Setup A

Data for 10,000 test takers were simulated for each replicate. Two uniformly distributed random variables, $D_1, D_2 \sim U(1, 5)$, were generated and used as background covariates. A treatment variable, Z, was generated as a sequence of 10,000 Bernoulli trials with probability of receiving test form X given by

$$\Pr(Z = 1 | \mathbf{D}) = e(\mathbf{D}) = (1 + \exp(3.5 - 0.6D_1 - 0.55D_2))^{-1}. \quad (10)$$

The propensity score was set so that about half of the test takers received form X and half of them received form Y.

The potential test scores for every test taker on the two simulated test forms were generated as

$$X = -6 + 4D_1 + 5D_2 + \epsilon_X \tag{11}$$

and

$$Y = -9 + 3D_1 + 6D_2 + \epsilon_Y, \tag{12}$$

where $\epsilon_X \sim N(2,\ 1.5)$ and $\epsilon_Y \sim N(0,\ 1)$. The covariates D_1 and D_2 act as a proxy for ability and differ in their distributions between the test groups, making the groups nonequivalent. The random noise terms ϵ_X and ϵ_Y represent the difference in difficulty between the forms. The expressions for X and Y were set so that the distributions of the scores mimicked real test data, with means 23 and 18 and standard deviations $\sqrt{683/12} \approx 7.54$ and $\sqrt{61} \approx 7.75$. Both X and Y were discretized by rounding each score to the nearest integer, and all scores above 40 were set to 40. This means that the discretized score variables are defined on the interval [0, 40]. The variables X and Y are thus thought of as scores from test forms containing 40 items each, scored as number of correct responses. The correlation of X and Y was, both before and after discretizing, approximately 0.94. From now on, X and Y always refer to the integer value scores.

The observed score was defined as

$$U = ZX + (1 - Z)Y. \tag{13}$$

After the potential and observed scores had been generated, the covariates D_1 and D_2 were discretized by categorizing them into five categories of approximately equal size delimited by percentiles of the scores. Two estimation models were set up for the propensity score, where one used the discretized covariates and one did not. In both cases, the propensity scores were estimated by the correctly specified logistic regression models, and the estimates were divided into 20 categories based on the percentiles. The number of categories was selected with the aim of balancing the covariate distribution in the two samples, as measured by the ASMD.

With the data generation process described, each test taker had a potential test score on both test forms given by Equations 11 and 12, an observed test score indicating what administration the test taker actually took given by Equation 13, an observed value on both covariates, and an observed value on the true propensity score given by Equation 10.

6.2. Simulation Design—Setup B

Ten thousand test takers were generated for each replicate. Two background random variables following the beta-binomial distribution were generated as

$D_1, D_2 \sim \text{Bin}(10,000, p)$, where $p = \text{Beta}(3,3)$. Furthermore, a treatment variable $Z \sim \text{Bern}(e(\mathbf{D}))$ was generated for every test taker, where

$$e(\mathbf{D}) = \Pr(Z = 1|\mathbf{D}) = (1 + \exp(3.7 - 0.25D_1 - 0.12D_2))^{-1}. \quad (14)$$

This meant that the test groups were of about equal size. The potential test scores were defined as:

$$X = -8 + D_1 + 1.5D_2 + \epsilon_X \quad (15)$$

and

$$Y = -5 + 1.2D_1 + 1.2D_2 + \epsilon_Y, \quad (16)$$

where $\epsilon_X \sim N(0,1)$ and $\epsilon_Y \sim N(2, 1.5)$. The distribution of the covariates D_1 and D_2 differs between the test groups, and the random noise terms ϵ_X and ϵ_Y represent the difficulty of the forms. In this way, the score distributions imitated real test data. The means of X and Y are 17 and 21, and the standard deviations are $\sqrt{997/22} \approx 6.73$ and $\sqrt{1827/44} \approx 6.44$. Both X and Y were discretized by rounding each score to the nearest integer, and every score exceeding 40 was set to 40. After the discretization, the score variables are defined on the interval [0, 40]. As for Setup A, the correlation of X and Y was both before and after discretizing approximately 0.94. The test takers' observed score was defined by Equation 13.

The covariates D_1 and D_2 were stratified into seven categories based on the percentiles. The propensity score defined in Equation 14 was estimated using both categorized and original continuous covariates for two separate models. Both models were correctly specified, and the estimated probabilities were categorized into 20 strata based on the percentiles.

The data generating process of Setup B yielded for every test taker a potential test score on the two forms by respective Equations 15 and 16, an observed test score defined by Equation 13, an observed value on both covariates, and an observed value on the true propensity score given by Equation 14.

6.3. Assessment of Precision

The three estimated equating transformations were evaluated in terms of bias, root mean squared error (RMSE) and standard error (SE), as defined in Wiberg and González (2016). Let $\hat{\varphi}_Y^{(g)}(x_i), i = 1, \ldots, 40$ denote the estimated equated test score of 40 possible scores for the gth replication using PSE NEC PS, CE NEC PS, or PSE NEC RAW COV. The methods are evaluated by the following measures:

$$\text{Bias}\,(\hat{\varphi}_Y(x_i)) = \frac{1}{1,000}\sum_{g=1}^{1,000}(\hat{\varphi}_Y^{(g)}(x_i) - \varphi_Y(x_i)),$$

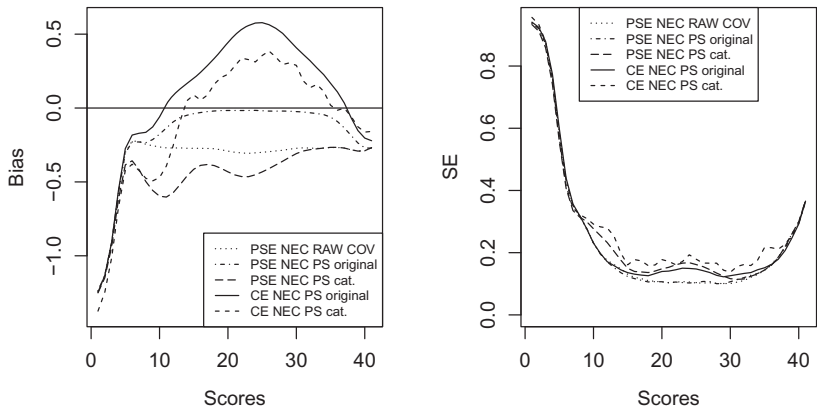

FIGURE 4. *The bias and* SE *for Setup A using PSE NEC RAW COV, PSE NEC PS, and CE NEC PS. Both categorized and original continuous covariates are considered for the propensity score methods. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating;* SE = *standard error.*

$$\text{RMSE}\left(\hat{\varphi}_Y(x_i)\right) = \sqrt{\frac{1}{1,000}\sum_{g=1}^{1,000}\left(\hat{\varphi}_Y^{(g)}(x_i) - \varphi_Y(x_i)\right)^2},$$

and

$$\text{SE}\left(\hat{\varphi}_Y(x_i)\right) = \sqrt{\frac{1}{1,000-1}\sum_{g=1}^{1,000}\left(\hat{\varphi}_Y^{(g)}(x_i) - \bar{\varphi}_Y^{(g)}\right)^2},$$

where $\varphi_Y(x)$ is the true equating transformation and $\bar{\varphi}_Y^{(g)} = \frac{1}{1,000}\sum_{g=1}^{1,000}\varphi_Y^{(g)}(x_i)$. In addition, the difference that matters (DTM; Dorans & Feigenbaum, 1994), defined as a difference larger than half a raw score point, was used. Note, however, that the DTM is presented not for the individual scores but for their summaries.

6.4. Results—Setup A

In Figure 4, the biases and SEs of the estimators are illustrated. From the left-hand panel, it is evident that all biases falls below -1 in the lower end of the score scale, and end up less than -0.5 in the upper end. The bias is smaller when the original continuous covariates are used instead of the categorized versions for the PSE NEC PS approach. For CE NEC PS, the bias is smaller when categorized covariates are used. For a large part of the score scale, PSE NEC PS using the

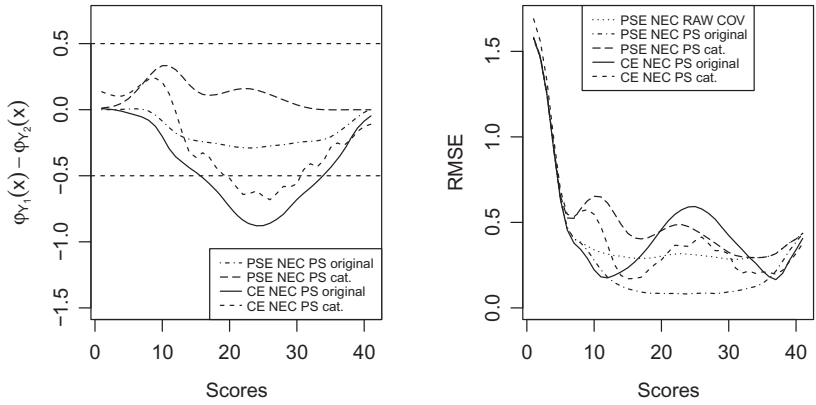

FIGURE 5. *Left panel: The difference between the mean of the 1,000 replicates for the propensity score equating transformations to the mean of the PSE NEC RAW COV transformation under Setup A. Dashed, horizontal lines represent the DTM. Both categorized and original continuous covariates are considered for the propensity score methods. Right panel: The RMSE under Setup A for PSE NEC RAW COV, PSE NEC PS, and CE NEC PS, the two latter considering both categorized and original covariates. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating; DTM = difference that matters; RMSE = root mean squared error.*

original covariates performs best. In fact, between scores 15 and 35, the PSE NEC PS using original continuous covariates is practically bias free. However, for the important upper score scale, CE NEC PS offers the smallest bias, both when considering categorized and original covariates. The right-hand panel illustrates only small differences between the estimators in terms of SE, with a maximum difference of only about 0.1, observed for PSE NEC PS with original covariates and CE NEC PS with categorized covariates. All SEs fall below 1, and the PSE NEC RAW COV shows the overall best performance along the score scale, and the CE NEC PS using categorized covariates the worst performance.

In the left-hand panel of Figure 5, the mean difference of each propensity score–based equating transformation to the PSE NEC RAW COV is presented, together with the DTM. It is only the two CE NEC PS approaches that deviate by more than half a score point from the equated scores using PSE NEC RAW COV. In the right-hand panel of Figure 5, the RMSE is presented. It varies between approximately 1.5 in the lower end and about 0.5 in the upper end. Similarly, as Figure 4 illustrated, PSE NEC PS using the original covariates shows the best performance for a large part of the score scale, but for the top 10 scores, the CE NEC PS approaches are the overall best equating transformations. The maximum difference in RMSE is seen for PSE NEC PS and CE NEC PS with original covariates, with a magnitude of about 0.5.

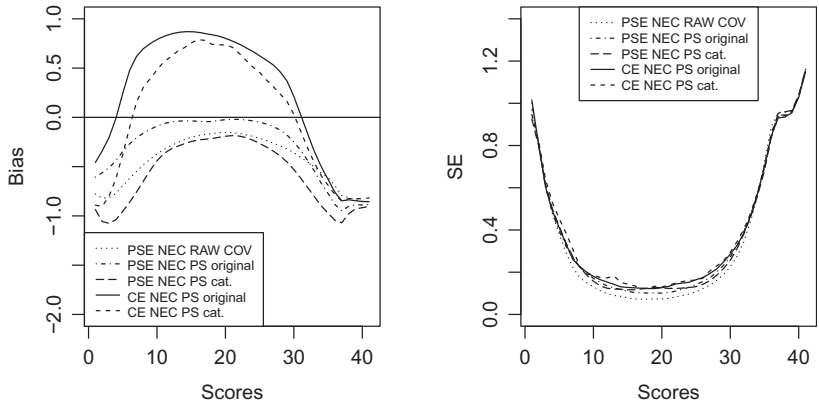

FIGURE 6. *Left panel: The bias of the 1,000 replications for PSE NEC RAW COV, PSE NEC PS, and CE NEC PS under Setup B. Both categorized and original covariates are considered for the propensity score methods. Right panel: The* SE *under Setup B for PSE NEC RAW COV, PSE NEC PS, and CE NEC PS. Both categorized and original covariates are considered for the propensity score methods. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating;* SE = *standard error.*

6.4. Results—Setup B

In Figure 6, the bias and SE are illustrated. In the left-hand panel, the biases fall within approximately ± 1 along the score scale. Similar to Setup A, PSE NEC PS using the original covariates performs best for a large part of the score scale, with a negligible bias in the score range 10 to 30. It is evident that PSE NEC RAW COV is better than the PSE NEC PS version that uses categorized covariates, which is different from Setup A. Again, the CE NEC PS approaches offer among the smallest biases for the top scores. The SEs in the right-hand panel of Figure 6 show that the equating estimators exhibit a very similar variation along the score scale. Both the bias and SE reflect the fact that the test forms were difficult, with few test takers getting top scores. This explains why the SE is approximately 1 for the top scores in Setup B and only about 0.3 for the same scores in Setup A.

In the left-hand panel of Figure 7, the mean difference to PSE NEC RAW COV is illustrated for each propensity score-based estimator. As in Setup A, it is only the CE NEC PS approaches that crosses the DTM bounds for some part of the score scale. In the right-hand panel, the RMSE is displayed. It varies between approximately ± 1.3 from the lower to the upper end of the score scale. It is similar to the RMSE of Setup A, meaning that PSE NEC PS using the original covariates has the overall best performance for the mid-scores and that there are only small differences for the top scores. As for Setup A, the maximum difference is seen for PSE NEC PS and CE NEC PS with original covariates, with a magnitude of about 0.75.

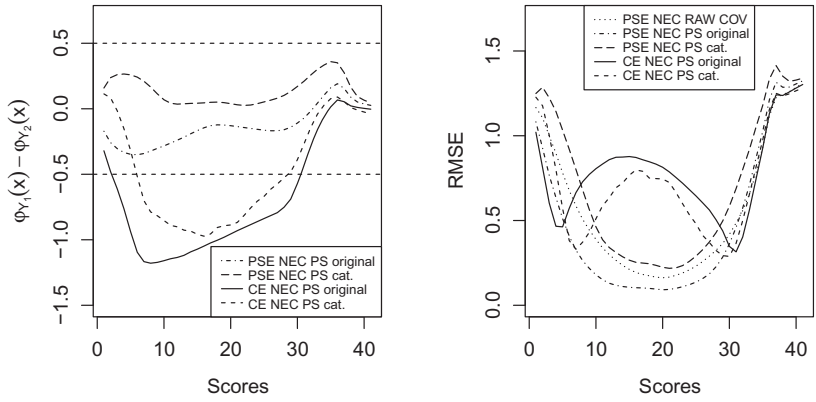

FIGURE 7. *Left panel: The difference between the mean of the 1,000 replicates for the propensity score equating transformations to the mean of the PSE NEC RAW COV transformation under Setup B. Dashed, horizontal lines represent the DTM. Both categorized and original covariates are considered for the propensity score methods. Right panel: The RMSE under Set-up B for PSE NEC RAW COV, PSE NEC PS, and CE NEC PS, the two latter considering both categorized and original covariates. PSE = poststratification equating; NEC = nonequivalent groups with covariates; CE = chained equating; DTM = difference that matters; RMSE = root mean squared error.*

7. Conclusions

We examined propensity scores in kernel equating, motivated by test situations where the test groups are nonequivalent and there is no anchor test available. This article also extends the NEC design by giving expressions for CE. In the empirical study, real admissions data from the SweSAT was used to exemplify how the PSE and CE approaches in the NEC design could be implemented using propensity scores. The results of the empirical study showed that all PSE approaches produced similar equated scores and that there was a clear difference from using the propensity scores in a CE approach. The results of the suggested methods were also clearly different from the results using an EG design and generally more similar to the results using a NEAT design, which is the design that the SweSAT uses today. In terms of SEE, the differences between the evaluated estimators were generally small, but the NEC design estimators all produced lower SEEs in the tails of the score distribution than the NEAT and EG design counterparts. Even though PSE NEC RAW COV resulted in the overall smallest SEEs, it was only by a small margin when considering the covariates within a propensity score instead. It is furthermore reasonable to believe that not all of the confounding covariates were observed which, if being available to condition on, would have improved the propensity score–based methods. However, it is hard to draw any reliable conclusions from only this data set.

This article also contributes with a simulation study, which evaluated and compared the three NEC design estimators for two different setups. It was evident that the information lost by discretizing the covariates affected the propensity score negatively for the PSE approach but positively for the CE approach. The estimators all performed similarly in terms of SE, and the RMSE made clear that they ordered themselves differently along the score scale. As an alternative to these measures one could, for example, rank the estimators according to how close they are to the true score and then average these ranks. Altogether, the differences were small between the categorized and original covariates of the novel equating approaches. This is important since some data sets only include discretized/categorized versions of the covariates, for example, the covariate Age in the SweSAT data.

In Setup B, discrete covariates were generated. The categorized covariates, as in Setup A, had a negative impact on the estimated propensity score for PSE NEC PS and a positive impact for CE NEC PS. It is also evident that the bias, SE, and RMSE were larger than they were in Setup A, especially for the top scores. This is likely due to sparse data, since two difficult forms were generated.

For both simulation setups, only the mean CE NEC PS function, for both categorized and original continuous covariates, deviated by more than half a unit from NEC PS RAW COV, which further indicates that the PSE approaches within the NEC design produce very similar equating results.

This article has clarified the underlying assumptions of the NEC design, expanded the ways that covariates can be used, and clearly stated the purpose set up for them through Assumptions A.1–A.3 and B.1–B.2. With much of the overall idea being borrowed from existing methods of the NEAT design, we have made both methodological and empirical comparisons to the NEAT design to point out differences. While the results of our study are promising, further studies are needed to determine the applicability of the NEC design. One such study could be to investigate how sets of covariates with different dependence structures affect the quality of the propensity score as a proxy for ability. Furthermore, the novel approaches presented are based on untestable assumptions. This makes them similar to the traditional PSE and CE approaches for nonequivalent groups using anchor scores. This does not mean that it is not of great importance to further evaluate the sensitivity of the results to the assumptions. Future studies should therefore investigate how to assess the sensitivity of all the untestable assumptions, including the model adopted here. It is also important to assess the implication of model misspecification of the propensity score, to compare different ways of estimating the propensity score, and to investigate alternative uses of the propensity score other than categorizing it. Lastly, the impact of presmoothing for the two propensity score approaches should be studied, preferably examining different real test data. Until these issues have been addressed, the NEAT design will act as the gold standard when equating nonequivalent groups.

Appendix A

Score Distributions of SweSAT

The test score distributions of the SweSAT data.

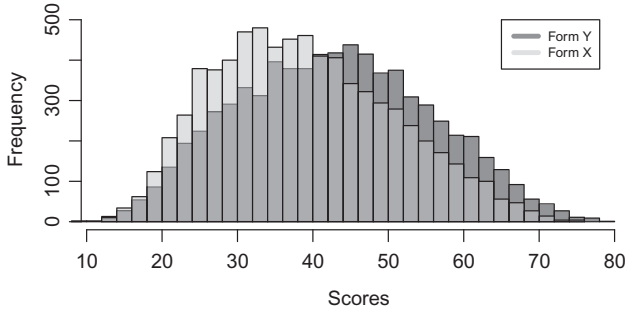

FIGURE A1. *The score distributions of the SweSAT data.*

Appendix B

Presmoothing and Estimation of the Score Probabilities

Let n_{Pjl} and n_{Qkl} be the number of test takers in the sample from populations P and Q with $X = x_j$ and $Y = y_k$ and with propensity score e_{Xl} and e_{Yl}, respectively. It is assumed that the vectors $\mathbf{n}_P = (n_{P11}, \ldots, n_{PJL})^t$ and $\mathbf{n}_Q = (n_{Q11}, \ldots, n_{QKL})^t$ are independent and that both follow a multinomial distribution, with $p_{jl} = \mathrm{P}(X = x_j, e_{Xl}(\mathbf{D}) = e_{Xl}(\mathbf{d})|P)$ and $q_{kl} = \mathrm{P}\Big(Y = y_k, e_{Yl}(\mathbf{D}) = e_{Yl}(\mathbf{d})|Q\Big)$ denoting the joint probabilities of the test scores and the categorized propensity scores for population P and Q, respectively. To reduce sampling error and get a more stable equating function, log-linear models can be used to fit the empirical joint distributions (Holland & Thayer, 2000), where all log-linear models should be based on model fit indices and empirical examinations.

SEE

Letting $\mathbf{P}$ be the $J \times L$ matrix of probabilities p_{jl} in population P and $\mathbf{Q}$ be the $K \times L$ matrix of probabilities q_{kl} in population Q, the Jacobian of the design function $\widehat{\mathbf{J}}_{\mathrm{DF}}$ is used to map probabilities in $\mathbf{P}$ and $\mathbf{Q}$ into $\boldsymbol{r}$ and $\boldsymbol{s}$. Matrix $\mathbf{C}$ is based on the covariance between the estimators of the probabilities in $\mathbf{P}$ and $\mathbf{Q}$. The mathematical definitions of these components for PSE NEC PS are similar to those given in Wiberg and Bränberg (2015), although propensity scores are used

instead of using covariates directly. Thus, the definition of the design function for PSE NEC PS is excluded here, although it can be sent upon request.

Let $v(\mathbf{P}) = (p_{11}, \ldots, p_{JL})^t$ and $v(\mathbf{Q}) = (q_{11}, \ldots, q_{KL})^t$ be vectorized versions of $\mathbf{P}$ and $\mathbf{Q}$. Letting subscripts indicate the population in question, the design function for CE is in general defined as:

$$\begin{pmatrix} \boldsymbol{r}_P \\ \boldsymbol{t}_P \\ \boldsymbol{t}_Q \\ \boldsymbol{s}_Q \end{pmatrix} = \mathrm{DF}(\mathbf{P}, \mathbf{Q}) = \begin{pmatrix} \begin{pmatrix} \mathbf{M}_P \\ \mathbf{N}_P \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} \mathbf{M}_Q \\ \mathbf{N}_Q \end{pmatrix} \end{pmatrix} \begin{pmatrix} v(\mathbf{P}) \\ v(\mathbf{Q}) \end{pmatrix},$$

where $\mathbf{M}$ is the $(J \times KJ)$ matrix with K number of $J \times J$-identity matrices $\mathbf{I}_J$ as elements, and $\mathbf{N}$ is a $(K \times KJ)$ matrix containing zeros and ones. To determine the SEE for CE NEC PS, the two links in Equations B1 and B2 can be viewed as equating transformations from two single group designs, meaning that the Jacobian of the equating transformation, $\mathbf{J}_{\varphi_Y}(x)$, can be formed from two Jacobians and the four sets of probabilities ($\boldsymbol{r}_P, \boldsymbol{t}_P, \boldsymbol{t}_Q$, and $\boldsymbol{r}_Q$). This Jacobian can be expressed as:

$$\mathbf{J}_{\varphi_Y}(x) = (\varphi'_Y(e_{YI})\mathbf{J}_{\varphi_{e_{XI}}}(x), \mathbf{J}_{\varphi_Y}(e_{YI})),$$

where $\varphi'_Y(e_{YI})$ is the derivative of $\varphi_Y(e_{YI})$ with respect to e_{YI}, $\mathbf{J}_{e_{XI}}$ is the Jacobian of $\varphi_{e_{XI}}$, and $\mathbf{J}_{\varphi_Y}$ is the Jacobian of φ_Y.

We assume that $\mathbf{P}$ and $\mathbf{Q}$ are estimated independently using log-linear models and maximum likelihood, and instead of searching for one $\mathbf{C}$ matrix as in Equation 3, we should find the covariances $\mathbf{V}$ of each of the two links in Equations B1 and B2. We thus find the following covariances:

$$\mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{r}}_P \\ \hat{\boldsymbol{t}}_P \end{pmatrix} = \mathbf{V}_P\mathbf{V}_P^t \text{ and } \mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{s}}_Q \\ \hat{\boldsymbol{t}}_Q \end{pmatrix} = \mathbf{V}_Q\mathbf{V}_Q^t, \text{ where}$$

$$\mathbf{V}_P = \begin{pmatrix} \mathbf{M}_P \\ \mathbf{N}_Q \end{pmatrix}\mathbf{C}_P \text{ and } \mathbf{V}_Q = \begin{pmatrix} \mathbf{N}_Q \\ \mathbf{M}_Q \end{pmatrix}\mathbf{C}_Q.$$

To obtain the SEE for CE NEC PS, we redefine Equation 3 as:

$$\mathrm{SEE}_Y(x) = ||\mathbf{J}_{\varphi_Y}\mathbf{V}||,$$

where

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_P & 0 \\ 0 & \mathbf{V}_Q \end{pmatrix}.$$

Appendix C

Table C1 specifies the considered presmoothing models for the empirical study. The first row of the log-linear models specifies, for every scenario, the included terms for the X scores, and the second row the included terms for the Y scores. Only the highest power of the variables is explicitly stated, so, for example, X^3 means that both X and X^2 are included. Colons refer to interactions between variables. The bandwidths in the continuization step are also given for each model.

TABLE C1.
Log-linear Models for the Four Different Scenarios (S) as well as Optimal Bandwidths for CE ($h_P, h_Q, h_{e_{XI}}$, and $h_{e_{YI}}$)*, PSE* (h_X and h_Y)*, and EG Designs* (h_X and h_Y)

S. No.	Log-Linear Models	h_P	h_Q	$h_{e_{XI}}$	$h_{e_{YI}}$	h_X	h_Y
1.	$X^3, PS^5, X:PS, X^2:PS, X:PS^2$ $Y^3, PS^5, Y:PS, Y^2:PS, Y:PS^2$	.653	.673	.467	.487	.662	.663
2.	X^5, V^2, A, $G^2, X:V, X:G, X^2:V, X^2:G$ Y^5, V^2, A, $G^2, Y:V, Y:G, Y^2:V, Y^2:G$	—	—	—	—	.661	.663
3.	$X^6, A^4, X:A$ $Y^6, A^4, Y:A, Y^2:A, Y^2:A^2, Y:A^2$	.653	.673	.578	.593	.661	.662
4.	X^6 Y^6	—	—	—	—	.653	.673

Note. S1 = PSE NEC PS and CE NEC PS, S2 = PSE NEC RAW COV, S3 = NEAT design (both PSE and CE), S4 = EG. CE = chained equating; PSE = poststratification equating; NEC = nonequivalent groups with covariates; EG = equivalent groups; NEAT = nonequivalent groups with an anchor; PS = propensity score; V = verbal test score; A = age; G = gender.

Appendix D

Absolute Standardized Mean Difference

TABLE D1.
The Absolute Mean Standardized Difference for Each Stratum and Covariate

Stratum	Verb	Gender	Age
1	.252	.164	.258
2	.055	.314	.139
3	.014	.100	.040
4	.157	.005	.042
5	.189	.061	.007
6	.091	.027	.010
7	.386	.064	.118
8	.003	.118	.073
9	.224	.323	.080
10	.094	.369	.054

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Swedish Research Council Grant 2014-578.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235–267.

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating using the R 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*, 1202–1217.

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

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

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1990). *Equating achievement tests using samples matched on ability* (College Board Report 90–2). New York, NY: College Entrance Examination Board.

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

Häggström, J., & Wiberg, M. (2014). Optimal bandwidth selection in kernel equating. *Journal of Educational Measurement*, *51*, 201–211.

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

Hsu, T., Wu, K., Yu, J., & Lee, M. (2009). Exploring the feasibility of collateral information test equating. *International Journal of Testing*, *2*, 1–14.

INVALSI. (2013). *Rilevazioni nazionali sugli apprendimenti 2012-13* (Technical Report). Retrieved from www.invalsi.it/snvpn2013/rapporti/Rapporto_SNV_PN_2013_DEF_11_07_2013.pdf/

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, *3*, 23–39.

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Measurement in Education*, *25*, 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*, 73–95.

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

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). Princeton, NJ: Educational Testing Service.

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. Retrieved from http://ir.uiowa.edu/etd/875

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

R Core Development Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for statistical computing. Retrieved from http://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*, 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. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, *47*, 261–285.

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

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*, 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*, 15–32.

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

Wang, T., Lee, W.-C., Brennan, R. B., & Kolen, M. J. (2006). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement*, *32*, 632–651.

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

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

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS Research Report RR-93–04). Princeton, NJ: Educational Testing Service.

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.

Authors

GABRIEL WALLIN is a PhD student at the Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, SE-901 87 Umeå, Sweden; email: gabriel.wallin@umu.se. His research interests include psychometrics in general and test equating in particular.

MARIE WIBERG is a professor at the Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, SE-901 87 Umeå, Sweden; email: marie.wiberg@umu.se. Her research interests include psychometric models and methods, test equating, and international large-scale assessments.

Manuscript received December 2, 2016
First revision received May 20, 2018
Second revision received October 29, 2018
Accepted February 20, 2019