Kernel Equating Under the Non-Equivalent Groups With Covariates Design

Applied Psychological Measurement

2015, Vol. 39(5) 349–361

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614567939

apm.sagepub.com

Marie Wiberg¹ and Kenny Bränberg¹

Abstract

When equating two tests, the traditional approach is to use common test takers and/or common items. Here, the idea is to use variables correlated with the test scores (e.g., school grades and other test scores) as a substitute for common items in a non-equivalent groups with covariates (NEC) design. This is performed in the framework of kernel equating and with an extension of the method developed for post-stratification equating in the non-equivalent groups with anchor test design. Real data from a college admissions test were used to illustrate the use of the design. The equated scores from the NEC design were compared with equated scores from the equivalent group (EG) design, that is, equating with no covariates as well as with equated scores when a constructed anchor test was used. The results indicate that the NEC design can produce lower standard errors compared with an EG design. When covariates were used together with an anchor test, the smallest standard errors were obtained over a large range of test scores. The results obtained, that an EG design equating can be improved by adjusting for differences in test score distributions caused by differences in the distribution of covariates, are useful in practice because not all standardized tests have anchor tests.

Keywords

equipercenile equating, non-equivalent groups design, background variables, NEC design

To make fair comparisons between test takers taking different test forms, one must find the equating function between the different scales of the test forms. The traditional approach to estimating the equating function between test forms is to use common test takers or common items (Kolen & Brennan, 2004). The problem is that one does not always have access to common test takers or common items, but one might have access to equivalent test takers, that is, allowing for an equivalent group (EG) design, or have access to other information. The focus in this article is on a data collection design that is closely related to the non-equivalent groups with anchor test (NEAT) design. In the NEAT design, one of the test forms is administered to one sample, and the other test form is administered to another sample, usually from a different population. An anchor test is administered to both samples and is used to adjust for differences in ability. It is impossible to apply the NEAT design in every instance because not all test

¹Umeå University, Sweden

Corresponding Author:

Marie Wiberg, Department of Statistics, Umeå School of Business and Economics, Umeå University, SE-901 87 Umeå, Sweden.

Email: marie.wiberg@umu.se

programs and standardized tests have anchor tests. A similar problem occurs if a test program uses an anchor test only to examine trends over time and thus only gives the anchor test to a limited number of test takers. Both of these problems might be related to attempts to ensure test security, to the design of the test, or to administrative problems. However, access to meaningful covariates provides an alternative. In the absence of an anchor test, the authors propose the non-equivalent groups with covariates (NEC) design. The equatings from the NEC design will be compared with equatings from the EG design, as this design has previously been used in several large-scale assessment tests that lack anchor tests, for example, the American College Testing (ACT) (2007), Armed Services Vocational Aptitude Battery (ASVAB; Quenette, Nicewander, & Thomasson, 2006), and Swedish Scholastic Aptitude Test (SweSAT). For the SweSAT, although an EG design has been used, the EG assumption has been demonstrated to be problematic (Lyrén & Hambleton, 2011), thus indicating that one may have non-equivalent groups. In the NEC design, covariates are used, such as grades, previous test scores, or other standardized test scores, as a substitute for an anchor test. The fundamental assumption is that these covariates can explain the differences in ability between the groups.

Using variables other than scores on an anchor test in equating or in other forms of linking is not a new idea. Kolen (1990) and Livingston, Dorans, and Wright (1990) suggested using groups matched on variables other than ability. Cook, Eignor, and Schmitt (1990) examined the efficacy of using data from responses on selected questions from the Student Descriptive Questionnaire for matching. Wright and Dorans (1993) used a selection variable, defined as the variable or set of variables in which the subpopulations differ, both for matching of subpopulations and as an anchor. Liou (1998) suggested a model for estimating score distributions using incomplete data that use background information together with common items as possible predictors of sample-selection bias. Liou, Cheng, and Li (2001) used a “surrogate variable” for the anchor test score that was correlated with the test score as well as some background variables such as gender and school attended. Finally, Holland, Dorans, and Petersen (2007) raised the question of whether it makes sense to use supplemental information about the test takers and suggested that this is an important area for further research. The present work can be viewed as a small step in that direction.

The choice of covariates is important, and a good rule is to aim for variables that correlate with the test scores. In a non-equivalent groups design, the goal would be to use variables that can explain possible group differences. In the numerical illustration presented in this work, the authors used criterion-referenced school grades and scores from another standardized test as the covariates. The choice was determined partly by the availability of the data and partly by examining the correlation with possible covariates. Although previous studies have used, for example, gender and education level (Bränberg, Henriksson, Nyquist, & Wedman, 1990; Bränberg & Wiberg, 2011), these variables might be somewhat problematic to use and thus, the authors have explored the possibility of using other standardized test scores and grades.

The observed-score equating framework kernel equating (von Davier, 2013; von Davier, Holland, & Thayer, 2004) is used with emphasis on how the NEC design works in its five steps: presmoothing, estimation of score probabilities, continuization, equating, and calculating the standard error of equating. The primary purpose of this article is to use information on covariates within the framework of kernel equating and to propose the use of the kernel equating methods within the NEC design. A secondary purpose is to illustrate the kernel equating methods under the NEC design by equating two forms of a college admissions test and comparing the results with the EG design, the NEAT design, and a mixture of the NEAT and the NEC design.

This article is structured as follows. In the first part, the theory needed for performing kernel equating with the NEC design is described. In the second part, the NEC design is illustrated with

real data from a college admissions test. Finally, the results are discussed together with possible implications and areas for further research.

NEC Design

The authors assume that they have a sample of test takers from each of two populations P and Q and that for each test taker, they also have observations of other variables correlated with the test scores $\mathbf{X}$ and $\mathbf{Y}$. The idea is to use information from these covariates as a substitute for an anchor test to control for differences in ability between the two groups. The vector of covariates can consist of variables such as grades, other test scores, educational background, and so on. $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_D)^t$ is denoted as the column vector of D discrete covariates. Observations on $\mathbf{V}$ are denoted $\mathbf{v}_l$ for $l = 1, \dots, L$, where L is the number of all possible combinations of values for the covariates $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_D$. In the later numerical illustration in which two covariates with four levels each is used, $L = 16$. In practice, L cannot be too large because the number of observations for each combination of values of the covariates becomes smaller as L grows.

Presmoothing

In presmoothing, a statistical model is fitted to the empirical distribution obtained from the sampled data. It is assumed that much of the irregularities observed in the empirical distributions are due to sampling error, and the goal of presmoothing is to reduce this error. A problem with presmoothing is that even if the variance of the equating estimators is reduced, systematic error (bias) can be introduced if the models only have a few parameters. However, Moses and Holland (2007) used log-linear models in a NEAT design and concluded that using reasonably complex models (i.e., preserving at least three moments in the marginal distributions) probably produces negligibly biased equating functions in many situations. In this article, log-linear models—estimated by maximum likelihood—were fitted to the data because such a method is commonly used in the kernel method of test equating (von Davier et al., 2004). Let n_{pjl} and n_{qkl} be the number of test takers in the sample from populations P and Q, respectively, and observations on $\mathbf{X}$ and $\mathbf{Y}$ be denoted by x_j for $j = 1, \dots, J$ and y_k for $k = 1, \dots, K$, respectively. Thus, they have $\mathbf{X} = x_j$, $\mathbf{Y} = y_k$, and $\mathbf{V} = \mathbf{v}_l$. The vectors $\mathbf{n}_P = (n_{p11}, \dots, n_{pJL})^t$ and $\mathbf{n}_Q = (n_{q11}, \dots, n_{qKL})^t$ are assumed to be independent and that they each have a multinomial distribution. $p_{jl} = P(\mathbf{X} = x_j, \mathbf{V} = \mathbf{v}_l \mid P)$ and $q_{kl} = P(\mathbf{Y} = y_k, \mathbf{V} = \mathbf{v}_l \mid Q)$ are the joint probabilities, and following log-linear models are assumed for p_{jl} in population P and q_{kl} in population Q:

$$\log(p_{jl}) = \alpha_P + \mathbf{u}_{Pjl}^t \boldsymbol{\beta}_P, \text{ and } \log(q_{kl}) = \alpha_Q + \mathbf{u}_{Qkl}^t \boldsymbol{\beta}_Q, \quad (1)$$

where α_P and α_Q are normalizing constants to ensure that the probabilities sum to 1, $\mathbf{u}_{Pjl}$ and $\mathbf{u}_{Qkl}$ are vectors of known constants, and $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$ are vectors of unknown parameters. For a thorough review of the use of log-linear models to describe discrete test score distributions, see Holland and Thayer (2000).

Estimation of the Score Probabilities

The target population T is defined as a mixture of populations P and Q, that is, $T = wP + (1 - w)Q$, where $0 \leq w \leq 1$. Let r_j be the probability of a randomly selected individual in population T scoring x_j on Test X and s_k be the probability on Test Y. The score distributions of $\mathbf{X}$ and $\mathbf{Y}$ over T can be obtained from

$$r_j = P(\mathbf{X} = x_j \mid \mathbf{T}) = w r_{Pj} + (1 - w) r_{Qj}, \quad (2)$$

and

$$s_k = P(\mathbf{Y} = y_k \mid \mathbf{T}) = w s_{Pk} + (1 - w) s_{Qk}, \quad (3)$$

where $r_{Pj} = P(\mathbf{X} = x_j \mid \mathbf{P})$, $r_{Qj} = P(\mathbf{X} = x_j \mid \mathbf{Q})$, $s_{Pk} = P(\mathbf{Y} = y_k \mid \mathbf{P})$, and $s_{Qk} = P(\mathbf{Y} = y_k \mid \mathbf{Q})$ are the score probabilities of $\mathbf{X}$ and $\mathbf{Y}$ in the two populations $\mathbf{P}$ and $\mathbf{Q}$. The distribution of $\mathbf{X}$ in $\mathbf{P}$ and of $\mathbf{Y}$ in $\mathbf{Q}$ can easily be estimated directly through the estimators

$$\hat{r}_{Pj} = \sum_l \hat{p}_{jl} \text{ and } \hat{s}_{Qk} = \sum_l \hat{q}_{kl}. \quad (4)$$

The problem is that there are no data to directly estimate r_{Qj} and s_{Pk} . The proposed solution is to use the covariates as a substitute for an anchor test. It is assumed that the conditional distribution of $\mathbf{X}$ given $\mathbf{V}$ and the conditional distribution of $\mathbf{Y}$ given $\mathbf{V}$ are the same in both populations. This is basically the same assumption as when using a NEAT design and post-stratification equating (PSE), but covariates are used instead of an anchor test. With these two assumptions, the distribution of $\mathbf{X}$ in population $\mathbf{Q}$ and the distribution of $\mathbf{Y}$ in population $\mathbf{P}$ can be estimated by the following:

$$\hat{r}_{Qj} = \sum_l \left(\frac{\hat{p}_{jl}}{\sum_j \hat{p}_{jl}} \cdot \sum_k \hat{q}_{kl} \right) \text{ and } \hat{s}_{Pk} = \sum_l \left(\frac{\hat{q}_{kl}}{\sum_k \hat{q}_{kl}} \cdot \sum_j \hat{p}_{jl} \right). \quad (5)$$

Continuization

Because test scores are typically discrete, a Gaussian kernel is usually used in kernel equating to continuize the two estimated discrete cumulative distribution functions (cdf's; von Davier et al., 2004). Although it is possible to use other kernels—for example, uniform or logistic—it has been demonstrated that these usually only yield small differences in the equated scores (Lee & von Davier, 2011). Because of this, the Gaussian kernel was used in this article. This means that instead of the discrete variable $\mathbf{X}$, a continuous variable is used:

$$\mathbf{X}(h_X) = a_X(\mathbf{X} + h_X \mathbf{Z}) + (1 - a_X)\mu_X, \text{ where } a_X = \sqrt{\sigma_X^2 / (\sigma_X^2 + h_X^2)}, \quad (6)$$

and where the bandwidth $h_X > 0$ is a constant, $\mathbf{Z}$ is a standard normal variable independent of $\mathbf{X}$, and $\mu_X = \sum_j x_j r_j$ is the mean of $\mathbf{X}$ in population $\mathbf{T}$. The term a_X is a constant where σ_X^2 is the variance of $\mathbf{X}$ in population $\mathbf{T}$. The variable $\mathbf{X}(h_X)$ has the same mean and the same variance as $\mathbf{X}$. The Gaussian kernel smoothing of the distribution of $\mathbf{X}$ is given by

$$F_{h_X}(x) = P(\mathbf{X}(h_X) \leq x) = \sum_j r_j \Phi \left(\frac{x - a_X x_j - (1 - a_X)\mu_X}{a_X h_X} \right), \quad (7)$$

where $\Phi(\cdot)$ is the standard normal distribution function. The Gaussian kernel smoothing of the distribution of $\mathbf{Y}$ is, in a similar way, given by

$$G_{h_Y}(y) = P(\mathbf{Y}(h_Y) \leq y) = \sum_k s_k \Phi \left(\frac{y - a_Y y_k - (1 - a_Y)\mu_Y}{a_Y h_Y} \right). \quad (8)$$

The bandwidths in the kernel equating framework are usually selected by minimizing a penalty function (von Davier, 2013; von Davier et al., 2004), although other methods exist (Häggström & Wiberg, 2014). To find h_x , one must minimize the penalty function

$$\text{PEN}(h_x) = \sum_j \left(\hat{r}_j - \hat{f}_{h_x}(x_j) \right)^2 + \kappa \cdot \sum_j A_j, \quad (9)$$

where $\hat{f}_{h_x}$ is the estimated density function of $\mathbf{X}(h_x)$ and κ is a constant typically equal to 0 or 1 depending on whether the second penalty term is used. $A_j = 1$ if either of the following conditions is satisfied: (a) $\hat{f}'_{h_x}(x_j - v) > 0$ and $\hat{f}'_{h_x}(x_j + v) < 0$ or (b) $\hat{f}'_{h_x}(x_j - v) < 0$ and $\hat{f}'_{h_x}(x_j + v) > 0$, where $\hat{f}'_{h_x}(x_j)$ is the derivative of $\hat{f}_{h_x}(x_j)$ and v is a constant typically chosen to be 0.25 (Lee & von Davier, 2011; von Davier, 2013). Otherwise, $A_j = 0$. Either the whole penalty function is used or only the first term. The first term preserves the characteristics of the distribution, but it could undersmooth the data. The second term ensures the smoothness of the continuized distribution by adding a constant if the derivatives of the continuized density function have different signs within a narrow range around a test score value.

Equating

The estimator of the equipercntile kernel equating function for equating $\mathbf{X}$ to $\mathbf{Y}$ in population T is obtained from Equations 7 and 8 if the cdf's are continuous and strictly increasing

$$\hat{e}_Y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)). \quad (10)$$

Standard Error of Equating

The estimated equated score $\hat{e}_Y(x)$ is based on a sample; thus, there is a random error associated with this point estimator. An estimate of each estimator's standard deviation, the standard error of equating (SEE), can be computed using the δ -method (Kendall & Stuart, 1977), which is a commonly used method (Kolen & Brennan, 2004; von Davier et al., 2004). In the kernel equating framework, SEE is defined as

$$\text{SEE}_Y(x) = \left\| \hat{\mathbf{J}}_{e_Y} \hat{\mathbf{J}}_{DF} \mathbf{C} \right\|, \quad (11)$$

where $\| \cdot \|$ denotes the Euclidean norm of the vectors and $\mathbf{C}$ is a matrix that is based on the covariances between the estimators of the probabilities in $\mathbf{P}$ and $\mathbf{Q}$. $\hat{\mathbf{J}}_{e_Y}$ is the Jacobian of the equated score, and $\hat{\mathbf{J}}_{DF}$ is the Jacobian of the Design Function (DF), that is, a function that maps probabilities in $\mathbf{P}$ and $\mathbf{Q}$ into $\mathbf{r}$ and $\mathbf{s}$ where $\mathbf{r} = (r_1, \dots, r_J)^t$ and $\mathbf{s} = (s_1, \dots, s_K)^t$. To obtain the parts needed to calculate the SEE, one starts with two matrices $\mathbf{P}$ and $\mathbf{Q}$. $\mathbf{P}$ is the $J \times L$ matrix of probabilities p_{jl} in population P , and $\mathbf{Q}$ is the $K \times L$ matrix of probabilities q_{kl} in population Q . Note that each column in the matrices represents a combination of values of the covariates. Let $\mathbf{v}(\mathbf{P}) = (\mathbf{p}_{11}, \dots, \mathbf{p}_{JL})^t$ and $\mathbf{v}(\mathbf{Q}) = (\mathbf{q}_{11}, \dots, \mathbf{q}_{KL})^t$ be vectorized versions of $\mathbf{P}$ and $\mathbf{Q}$. These probabilities are estimated in the presmoothing step. Let t_{P_m} and t_{Q_m} be the sum of the m th column in matrices $\mathbf{P}$ and $\mathbf{Q}$, respectively. The j th element of vector $\mathbf{r}$ is then

$$r_j = w r_j + (1 - w) r_{Qj} = \sum_{m=1}^L \left[w + \frac{(1 - w) t_{Q_m}}{t_{P_m}} \right] \cdot p_{jm},$$

and s_k is obtained similarly. Thus, the DF for the NEC design is defined as

$$\begin{pmatrix} \mathbf{r} \\ \mathbf{s} \end{pmatrix} = \text{DF}(\mathbf{P}, \mathbf{Q}; w) \begin{pmatrix} \sum_m \left[w + \frac{(1-w)t_{Qm}}{t_{Pm}} \right] \mathbf{p}_m \\ \sum_m \left[(1-w) + \frac{wt_{Pm}}{t_{Qm}} \right] \mathbf{q}_m \end{pmatrix}, \quad (12)$$

where $\mathbf{p}_m$ and $\mathbf{q}_m$ are the m th column of $\mathbf{P}$ and $\mathbf{Q}$, respectively, and the summation is over all possible combinations of values of the covariates. The Jacobian of the DF is obtained by using the previously defined vectorized versions of $\mathbf{P}$ and $\mathbf{Q}$ ($\mathbf{v}(\mathbf{P})$ and $\mathbf{v}(\mathbf{Q})$, respectively)

$$\mathbf{J}_{DF} = \begin{pmatrix} \frac{\partial \mathbf{r}}{\partial \mathbf{v}(\mathbf{P})} & \frac{\partial \mathbf{r}}{\partial \mathbf{v}(\mathbf{Q})} \\ \frac{\partial \mathbf{s}}{\partial \mathbf{v}(\mathbf{P})} & \frac{\partial \mathbf{s}}{\partial \mathbf{v}(\mathbf{Q})} \end{pmatrix}. \quad (13)$$

Next, the Jacobian of the equating function is needed. For F and G (the two smooth cdf's), the Jacobian $\mathbf{J}_{e_Y}$ is defined as the $I \times (J \times K)$ vector

$$\mathbf{J}_{e_Y} = \left(\frac{\partial e_Y}{\partial r}, \frac{\partial e_Y}{\partial s} \right) = \frac{1}{G'} \left(\frac{\partial F}{\partial r}, -\frac{\partial G}{\partial s} \right), \quad (14)$$

where G' is the density of G evaluated at $e_Y(x)$. The estimate of the Jacobian is obtained by substituting the cdf's $\hat{F}_{h_X}$ and $\hat{G}_{h_Y}$ for F and G . Finally, to obtain the last part—the $\mathbf{C}$ matrix—the authors assume that $\mathbf{P}$ and $\mathbf{Q}$ are estimated independently using log-linear models and maximum likelihood. This leads to

$$\text{cov} \begin{pmatrix} \mathbf{v}(\hat{\mathbf{P}}) \\ \mathbf{v}(\hat{\mathbf{Q}}) \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{\mathbf{v}(\hat{\mathbf{P}})} \mathbf{C}_{\mathbf{v}(\hat{\mathbf{P}})}^t & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{v}(\hat{\mathbf{Q}})} \mathbf{C}_{\mathbf{v}(\hat{\mathbf{Q}})}^t \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{\mathbf{v}(\hat{\mathbf{P}})} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{v}(\hat{\mathbf{Q}})} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{\mathbf{v}(\hat{\mathbf{P}})}^t & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{v}(\hat{\mathbf{Q}})}^t \end{pmatrix} = \mathbf{C} \mathbf{C}^t, \quad (15)$$

where $\mathbf{C}$ is obtained from factorization of the covariance matrix. The size of the covariance between r and s is assumed to be small to simplify the calculation of SEE. The $\mathbf{C}$ matrix thus only contains the variances of the estimators in Equation 7. For more explicit details on computation of the $\mathbf{C}$ -matrices $\mathbf{C}_{\mathbf{v}(\hat{\mathbf{P}})}$ and $\mathbf{C}_{\mathbf{v}(\hat{\mathbf{Q}})}$ in the general kernel equating framework, see von

Davies et al. (2004) or Holland and Thayer (2000). Note that if it is assumed that $\begin{pmatrix} \mathbf{v}(\hat{\mathbf{P}}) \\ \mathbf{v}(\hat{\mathbf{Q}}) \end{pmatrix}$ is normally distributed with mean $\begin{pmatrix} \mathbf{v}(\mathbf{P}) \\ \mathbf{v}(\mathbf{Q}) \end{pmatrix}$ and variance $\mathbf{C} \mathbf{C}^t$, then for each x , the estimator of the equated score, $\hat{e}_Y(x)$, given by Equation 10 is also normally distributed.

Numerical Illustration

To illustrate the NEC design in kernel equating, data from the Swedish Scholastic Aptitude Test (SweSAT) college admissions test were used. The SweSAT is a paper-and-pencil test that traditionally has not administer an anchor test to a larger group due to test security issues. The SweSAT consists of a quantitative section and a verbal section. Both sections nowadays consist of 80 multiple-choice items, and the two sections are equated separately. The SweSAT is given twice a year (spring and fall) and has traditionally been equated using a matching procedure in which covariates play a major role to assure that similar groups are compared in the equating process. For a description of the equating methods that have been used for the SweSAT, see Lyrén and Hambleton (2011). The five-step kernel equating procedure was carried out in R (R

Core Development Team, 2013) using the kernel equating package *kequate* (Andersson, Bränberg, & Wiberg, 2013). The specific code can be obtained from the authors on request.

Procedure

The authors wanted to equate two consecutive test forms, a spring administration denoted as Test X and the previous fall administration denoted as Test Y, of the quantitative section. To examine the NEC design and allow for comparisons with both an EG design and a NEAT design, the authors used test takers who took both the spring and fall administrations and constructed a 24-item anchor test by selecting 12 items from both of the two administrations. The number of observations for test takers who took both test forms was 14,644. This group was divided in half and made sure that the distributions of covariates differed in each of the samples in a similar way as is common in the population of test takers who take the fall and the spring administration. Thus, a sample of 7,322 was used for each of the X and Y tests. The covariates were criterion-referenced grades and categorized test scores from the verbal section because these scores are known to correlate with the quantitative test score. The notation V_1 was used for the categorized grade—a variable with four possible values—and V_2 was used for the categorized verbal test scores. V_1 and V_2 both have 1 as the lowest value and 4 as the highest value, that is, $L=16$. The grades, which ranged from 0 to 320, were categorized as follows: Grade 1 = 0-225, Grade 2 = 226-255, Grade 3 = 256-290, and Grade 4 = 291-320. The verbal test score was categorized as Vtest 1 = 0-30, Vtest 2 = 31-40, Vtest 3 = 41-50, and Vtest 4 = 51-80. Descriptive statistics of the samples are given in Table 1. From Table 1, it is clear that categories with higher grades had higher quantitative test scores and that categories with higher verbal scores had higher quantitative test scores. The correlations were very strong between the two tests and the anchor. The correlation between the test scores and the covariates was positive and strong. The values obtained using the original values (Grade and Vtest) were similar to those obtained using the categorized values (GradeC and VtestC). The equatings using the NEC design were compared with equipercentile equating with EG design, NEAT Chain equating, NEAT PSE, and a mixture of NEC design with the two covariates and an anchor test (NEATNEC design) in the two groups. For comparison, an equipercentile equating with all 14,644 test takers taking the X and Y tests in the EG design (EGall) was also performed. It is important to note that one might argue that an EG design is not suitable here because the groups might be non-equivalent. However, the SweSAT has traditionally used an EG design due to the lack of an anchor test (Lyrén & Hambleton, 2011).

Numerical Illustration in the NEC Design

Let p_{jl} be the joint probability of the outcome and the covariates in population P and q_{kl} be the joint probability in population Q. These probabilities can be organized in matrices **P** and **Q**. Each column is a combination of the values of the covariates. It is assumed that the probabilities in **P** and **Q** can be described by log-linear models of the form presented in Equation 1.

Presmoothing

The models were evaluated with the goodness of fit statistics deviance and the Akaike Information Criterion. The principle of parsimony was followed in the model selection process. As a rule of thumb, the chi-square distribution for the deviance was used and a model was concluded to fit the data adequately if the p value was greater than .05. The best model was chosen for each data collection design. For both EGall and the sample EG, the log-linear models with

Table 1. Descriptive Statistics of the Samples Used in the Real Data Example With Four Grade Classes and Four Verbal Test Score Classes: Number of Test Takers (*n*), Mean, Standard Deviation, and the Correlation Between the Tests and the Anchor and Covariates.

Test scores	Sample from P			Sample from Q		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Grade = 1	2,105	33.37	10.21	1,416	35.55	11.72
Grade = 2	1,859	37.36	10.53	1,760	40.66	11.86
Grade = 3	1,991	41.96	11.29	2,306	44.66	11.78
Grade = 4	1,367	47.42	10.68	1,840	50.15	10.99
Vtest = 1	2,000	32.48	9.67	1,126	33.52	10.58
Vtest = 2	2,400	37.98	10.58	1,803	40.50	11.27
Vtest = 3	1,700	42.41	10.80	2,085	45.11	11.89
Vtest = 4	1,222	48.99	10.63	2,308	48.68	11.91
Correlation	Test Y	Anchor	Grade	GradeC	Vtest	VtestC
Test X	.86	.81	.43	.41	.47	.45
Test Y	1.00	.81	.46	.45	.46	.45

Note. GradeC and VtestC refer to categorized grades and verbal test scores, respectively.

six univariate moments demonstrated adequate fit in both population P and Q. For the following models, only the highest polynomial degree for the univariate moments are mentioned, although the lower polynomials are also included in the models, for example, X^3, A^2 means that X^1, X^2, X^3, A^1, A^2 are included in the given model. In the NEAT design, Test X demonstrated adequate fit for model (X^6, A^4, XA), and Test Y demonstrated adequate fit for model ($Y^6, A^4, YA, Y^2A, Y^2A^2, Y:A^2$). For the NEC design, Test X demonstrated adequate fit for model ($X^6, grade^3, vtest^3, X:vtest, X:verb, grade : vtest, X^2:grade, X^2:vtest, X^2:vtest$), and Test Y demonstrated adequate fit for model ($Y^6, grade^3, vtest^3, Y:grade, Y:vtest, grade:vtest, Y^2:grade, Y^2:vtest, Y^2:grade^2, Y^2:vtest^2$). For the NEATNEC design, the model ($X^6, A^5, vtest^3, X : A, X^2:A, X:A^2, X^2:A^2, X:vtest, A:vtest$) demonstrated adequate fit of the observed distribution ($p > .05$) in both population P and population Q.

Estimation of the Score Probabilities

The marginal probabilities $\hat{r}_{Pj}$, $\hat{s}_{Qk}$, $\hat{r}_{Qj}$, and $\hat{s}_{Pk}$ were obtained from Equations 4 and 5. The weight w was set to 0.50 in Equations 2 and 3. Although it is possible to give each population a weight according to its relative size, in this article, the populations are given the same weight to facilitate comparisons with the EG design. For a thorough discussion of the choice of weights, refer to Kolen and Brennan (2004, Section 4.1.5).

Continuization

The estimated cdf's were obtained from Equations 7 and 8. To use the Gaussian kernel smoothing in practice, the authors used the estimated distributions $\hat{r}_j$ and $\hat{s}_k$, the bandwidths h_X and h_Y , and estimates of the means and variances of **X** and **Y** in population T. The bandwidths were computed by minimizing the penalty function in Equation 9. In this case, the distributions were unimodal without “teeth” or “gaps,” so the constant κ was set to 0 in Equation 9. The obtained bandwidths were $h_X = 0.661$ and $h_Y = 0.662$ for NEC, $h_X = 0.661$ and $h_Y = 0.662$ for NEAT PSE,

Figure 1. Difference between the equated scores and the number-correct scores for the NEC, NEAT PSE, NEAT Chain, and EG designs for the samples, for the entire population (EGall), and for a NEATNEC. Note. NEC = non-equivalent groups with covariates; NEAT = non-equivalent groups with anchor test; PSE = post-stratification equating; EG = equivalent group; NEATNEC = NEC design with anchor test.

$h_X = 0.653$ and $h_Y = 0.673$ for NEAT Chain, $h_X = 0.651$ and $h_Y = 0.671$ for EG, $h_X = 0.662$ and $h_Y = 0.660$ for NEATNEC, and $h_X = 0.667$ and $h_Y = 0.653$ for the whole population EGall.

Equating and Calculating the Standard Error of Equating

The differences between the equated scores and the number-correct scores for different data collection designs are displayed in Figure 1. As a comparison, the results are displayed if no covariates were used and the two test taker groups were treated as statistically equivalent (EG design); the results if a regular NEAT design was used either with chain equating or PSE; and the results if a NEC design was used with the two covariates and an anchor test (NEATNEC design). These results indicated that some of the differences between the two samples could be explained by differences in the distribution of the covariates. The difference between equated scores and number-correct scores were smaller when the NEC design was used than when the EG design was used, indicating the possible value of using covariates in equating. The NEAT Chain and NEAT PSE had smaller differences than the NEC design. In other words, if there was access to an anchor test, the difference was smaller. These equatings can be compared with the “true” equating when the whole group took both tests (EGall). It is evident that using the EG design gave different equatings compared with those of the NEAT, NEC, and NEATNEC designs, all of which gave somewhat similar results.

The calculations needed for SEE are, in principle, straightforward but, in practice, are rather tedious because they involve large matrices. In Figure 2, the SEEs for the different data collection designs are displayed. In general, the SEE is large at the lower end of the scale and also somewhat larger at the upper end of the scale. This is expected considering the fact that there

Figure 2. Standard error of equating (SEE) for NEC, NEAT PSE, NEAT Chain, EG and NEATNEC designs. Note. The left figure displays the whole test score range, whereas the right figure displays a limited score scale without the endpoints. NEC = non-equivalent groups with covariates; NEAT = non-equivalent groups with anchor test; PSE = post-stratification equating; EG = equivalent group; NEATNEC = NEC design with anchor test.

were few test takers with very low results and very few with the highest result. The comparison of SEE between the NEC, NEAT Chain, NEAT PSE, and the NEATNEC design exhibits similar SEE values for all these measures, as shown in the left part of Figure 2. The right part of Figure 2 indicates that the NEC design has a smaller SEE than the EG design and a larger SEE than the NEAT designs. It is also interesting to note that using both covariates and the anchor test (NEATNEC design) resulted in the smallest SEE over a large range of test scores.

Discussion

This article presents a way of incorporating covariates in the framework of kernel equating through the NEC design. Because the samples in the NEC design are from two populations of test takers, there might be systematic differences in ability between the groups. An anchor test, as used in the NEAT design, is usually used to adjust for these differences. One of the methods for equating two test forms with the NEAT design is PSE, which has the basic assumption that the distribution of test scores, conditional on the anchor test score, is the same in both populations. A similar approach was used here; however, instead of an anchor test, it was proposed to use other variables—such as grades and other test scores—that correlate with the test scores. It is preferable to use anchor tests, but there are situations where there is no access to an anchor test. In such situations, a possible strategy is to use an EG design although the groups might not be equivalent as the EG assumption might be violated (Lyrén & Hambleton, 2011). In other situations, the anchor test might only have been administered to a limited number of test takers, thus providing only limited information. In the example presented here, the NEC design displayed smaller SEE and smaller differences between the number-correct scores and equated scores than when the covariates were ignored and an EG design was used on the same sample. It was noticeable that having access to an anchor test resulted in even smaller differences between the number-correct scores and the equated scores as well as smaller standard errors. If there was access to both an anchor test and covariates, it resulted in the smallest SEE over a large range of test scores. Using real-life data is always limiting because the true equating is not known, and here, an EG design was used in which all test takers taking both tests were used as a reference. The example in this article was used for illustrative purposes, and in the future, more research, especially using simulations for which the true equating is known, should be conducted to examine the NEC design within the kernel equating framework. Such research could answer the question of how much one can benefit from using a NEC design compared with using an EG design or possibly a mixture of the NEAT and the NEC designs.

The critical assumption in the NEC design is that the conditional distribution of the test scores, given the covariates, is the same in both populations. If the test forms are administered at different time points, and if the relationships between the test scores and the covariates change over time, this assumption might not be valid. An example of this would be if the ability, measured by the test, increases in each subpopulation defined by the values of the covariates. If the change is slow and if the time interval between the administrations of the test form is not too long, the bias will likely be small. However, if the change is rapid and/or if the time interval between the equated test forms is long, such bias can definitely become a problem. A possible solution is to include a variable in the model that explains the changes over time. In the absence of scores on an anchor test, criterion-referenced grades—such as the grades used in Swedish schools—could perhaps be such a variable. If other standardized tests are given, the scores from those tests could also be used. The general motivation for the proposed approach has been the existence of standardized tests that lack anchor tests or for which there are only a limited number of test takers who take the anchor test in addition to non-equivalent groups of test takers that possibly change over time. The results of the analysis indicated that if there is

no access to an anchor test and if the alternative is an EG design, incorporating information from covariates is useful if there exist differences in the test score distributions caused by differences in the distribution of covariates. It should also be noted that even if there is an anchor test, covariates can still be used as it may lead to lower SEE over a large range of test scores. The size of the difference in equating depends on how well the covariates predict test scores and how well background variables “explain” differences in test scores. Choosing the “right” covariates—variables that explain the differences between groups—could be an important area for future research.

Two problems that need to be highlighted when selecting the appropriate covariates are the number of categories within the covariates and how to construct the categories when the covariates are continuous. If one only uses a single covariate that correlates highly with the test score, one has a situation similar to the use of an anchor test. As one adds more covariates, one obtains a rapid increase in the number of categories. This means that the risk of having sparse data in some cells is increased. This is, of course, also dependent on how many test takers take the tests. To avoid sparse data in some cells, it is better to limit the amount of covariates unless they have very few categories. If there are continuous variables, it might be difficult to create reasonably delimited categories to avoid having too many categories.

In this study, two covariates are used, and although their correlations were not as high as the anchor test scores, they still provided useful information in the equatings. If even better covariates could be found, it is likely that the equating would become better and would more closely resemble the results obtained with the NEAT design. It should be noted that if the correlation had been even lower, one could have questioned the choice of adjusting the PSE for the use of covariates. Thus, an obvious area for further research would be to examine the use of chain equating because this is less influenced by correlations.

In the numerical illustration, relatively large sample sizes were used. One could argue that the presmoothing step is unnecessary with these large sample sizes, which might be true. However, the authors used the data primarily as an illustration of a general method. An additional, secondary reason for using presmoothing in this case is to obtain non-zero estimates of the probabilities close to the end of the score scale. Although there are large samples, there are many relative frequencies equal to zero close to the lower end of the score scale. Although these probabilities are small, they are not equal to zero.

The focus in this article has been on a non-equivalent groups design with no anchor test, but covariates can be used in other designs as well. In fact, using more information, such as variables correlated with the test scores, can reduce random error and increase the efficiency of the estimators of the equating parameters (Bränberg & Wiberg, 2011). Collecting this information is sometimes not too difficult and is also performed routinely in some testing programs (such as the SweSAT). How to use this information in an effective way is an area for future research. Finally, although using an anchor test is more efficient than using only covariates (Bränberg & Wiberg, 2011), there are standardized tests that do not administer an anchor test, or the anchor test is only given to very few test takers. Thus, other possibilities have to be explored to control for systematic differences. When using an EG design, it is possible that the EG assumption is violated (Lyrén & Hambleton, 2011). The fact that it is possible to use covariates in kernel equating and that it is possible to improve an EG design equating by adjusting for systematic differences caused by differences in the distribution of covariates is intriguing.

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: The research was partly financed by the Swedish Research Council.

References

- ACT. (2007). *ACT technical manual*. Iowa City, IA: Author.
- 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.
- 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 No. 90-2). New York, NY: College Entrance Examination Board.
- Häggström, J., & Wiberg, M. (2014). Optimal bandwidth in observed-score kernel equating. *Journal of Educational Measurement*, 51, 201-211.
- Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 169-203). Oxford, UK: Elsevier.
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25, 133-183.
- Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). New York, NY: Macmillan.
- Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3, 23-39.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Lee, Y.-H., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking* (Chap. 10, pp. 159-173). New York, NY: Springer.
- Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica*, 8, 669-690.
- Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, 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.
- Lyrén, P.-E., & Hambleton, R. K. (2011). Consequences of violated equating assumptions using the equivalent groups design. *International Journal of Testing*, 11, 308-323.
- Moses, T., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (ETS Research Report No. RR-07-15). Princeton, NJ: Educational Testing Service.
- 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. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for statistical computing. Available from <http://www.R-project.org/>
- von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, 78, 605-623.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.
- Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton, NJ: Educational Testing Service.