Research Report

An Equipercentile Version of the Levine Linear Observed-Score Equating Function Using the Methods of Kernel Equating

Alina A. von Davier

Stephanie Fournier-Zajac

Paul W. Holland

Research & Development

April 2007
RR-07-14

An Equipercentile Version of the Levine Linear Observed-Score Equating Function Using the Methods of Kernel Equating

Alina A. von Davier, Stephanie Fournier-Zajac, and Paul W. Holland

ETS, Princeton, NJ

April 2007

An Equipercentile Version of the Levine Linear Observed-Score Equating Function Using the Methods of Kernel Equating

Alina A. von Davier, Stephanie Fournier-Zajac, and Paul W. Holland

ETS, Princeton, NJ

April 2007

Abstract

In the nonequivalent groups with anchor test (NEAT) design, there are several ways to use the information provided by the anchor in the equating process. One of the NEAT-design equating methods is the linear observed-score Levine method (Kolen & Brennan, 2004). It is based on a classical test theory model of the true scores on the test forms to be equated and on the anchor test (Levine, 1955). The Levine linear method does not yet have an equipercentile analogue, and no version of kernel equating (KE), introduced in von Davier, Holland, and Thayer (2004b), approximates the Levine linear method. Nevertheless, the Levine observed-score equating method is often computed in practical applications for comparison purposes because it is sometimes more accurate than other linear equating methods (Petersen, Marco, & Stewart, 1982). In situations when a linear equating function is not satisfactory, an equipercentile version of the Levine function may be desirable. This paper proposes a general method for constructing *hybrid* equating functions that combine linear and nonlinear equating functions in a systematic way that preserves the symmetry required of equating functions (Dorans & Holland, 2000). The general method is then applied to combine the linear Levine observed-score equating function with a nonlinear equipercentile equating function derived using the poststratification equating (PSE) assumptions within the KE framework. An easily computed approximation to the resulting PSE-Levine equipercentile equating function is illustrated on data from a special study and compared to the results from the traditional equating functions. The special data set includes a criterion equating function, and the closeness of the PSE-Levine function to this criterion indicates that such hybrid equating functions may be useful in practice.

Key words: Linear equating, hybrid equating functions, kernel equating (KE), nonequivalent groups with anchor test (NEAT) design, anchor test design, common items, poststratification

Acknowledgments

The work that led to this paper was collaborative in every respect, and the order of authorship is alphabetical. The authors thank Dan Eignor and Wendy Yen for their comments on the previous version of this paper.

Introduction

This paper presents a way to create an equipercentile version of the Levine linear observed-score equating method. It uses ideas from von Davier, Holland, and Thayer (2004b) and exploits the general structure of the kernel equating (KE) framework for test equating. We present a general theoretical proposal and some empirical results that are derived under stronger assumptions than the general theory. With modest changes, available software can be used to implement the more general theoretical approach. This paper focuses on observed-score equating methods and does not address any of the item response theory methods for equating tests. See, for example, Kolen and Brennan (2004) for details on item response theory equating methods.

In the nonequivalent groups with anchor test (NEAT) design (also called the *common item* or *anchor test* design), two test scores are equated, X and Y, which are taken, respectively, by two samples of examinees each drawn from a different population, P or Q. When the equating function goes from X scores to Y scores, X is often called the score on the "new" form and Y the score on the "old" form. In the NEAT design, it is not assumed that P and Q are similar in any way. To deal with this, there is also the score on a set of *common items*, A, that is available for the examinees in both samples. This data collection arrangement is shown in the design table (von Davier et al., 2004b), illustrated in Table 1. In Table 1, the checkmark, ✓, denotes that examinees in the sample indicated by the rows have scores on the test indicated by the columns.

Table 1

The Design Table for the Nonequivalent Groups With Anchor Test (NEAT) Design

	X	A	Y
P	✓	✓	
Q		✓	✓

In the framework of the observed-score equating methods for the NEAT design, there are three fundamentally different ways of using the information provided by the anchor score, A, to equate the scores of X to those of Y. One method uses A as a conditioning variable (or covariate). In this method, the conditional distributions of X given A and of Y given A are weighted by a distribution for A to estimate the score distributions (or their first two moments) for X and Y in a hypothetical target population, T. T is an example of a *synthetic population*, a concept introduced in Braun and Holland (1982), and denoted there as $T = wP + (1 - w)Q$. The fraction, w, is the proportion of T that comes from P. This use of A is reminiscent of poststratification in survey research, and we follow von Davier et al. (2004a, 2004b) in referring to methods based on this approach as *poststratification equating* (PSE).

The PSE methods include both linear and equipercentile methods. Examples of linear PSE methods include the Tucker method (Kolen & Brennan, 2004), the Braun-Holland method (Braun & Holland, 1982; Kolen & Brennan, 2004) and the PSE linear method of KE (von Davier et al., 2004b). The PSE equipercentile methods include both frequency estimation (Kolen & Brennan, 2004) and the KE method of equipercentile PSE (von Davier et al., 2004b).

A second way to use A is as the middle link in a chain of linking relationships—X to A and A to Y. We will refer to equating methods based on this approach as *chain equating* (CE). An important difference between PSE and CE is that in the former there is an explicit target population, T, whereas in the latter T plays no *explicit* role. However, von Davier et al. (2004a, 2004b) showed that in order for CE to produce bona fide observed-score equating functions, certain assumptions must hold that involve an implicit synthetic population, T.

The CE approach also includes both linear and equipercentile methods. Examples of CE linear methods include chain linear equating (Angoff, 1971/1984; Livingston, 2004) and the KE method of linear CE (von Davier et al., 2004b). The CE equipercentile methods include chain equipercentile equating (Angoff, 1971/1984; Livingston, 2004) and the KE method of equipercentile CE (von Davier et al., 2004b).

The third use of A in the NEAT design is the Levine linear method (Kolen & Brennan, 2004; Levine, 1955). This method uses a classical test theory model for X, Y, and A to estimate the means and variances of X and Y on the target population from PSE, T. These four moments are sufficient to estimate a linear equating function, defined below in (5).

In this paper, we propose a general way to create equipercentile versions of the Levine linear method using the methods of KE. An approximate version of this approach is illustrated with data from a special study.

Review of the Levine Observed-Score Linear Method

The linear Levine observed-score equating was originally proposed by Levine (1955) and further developed in Kolen and Brennan (2004).

We assume a classical test theory model for X, Y and A, as shown in (1):

$$X = \tau_X + \varepsilon_X,\ Y = \tau_Y + \varepsilon_Y, \text{ and } A = \tau_A + \varepsilon_A, \qquad (1)$$

where the error terms, ε_X, ε_Y, and ε_A, have zero expected values and are uncorrelated with each other and with the true scores, τ_X, τ_Y, and τ_A, over any target population of the synthetic form, $T = wP + (1 - w)\, Q$ and for any choice of $0 \le w \le 1$. From (1), the basic equations in (2) follow for any T of this form:

$$\mu_{XT} = \mathrm{E}(Y|\ T) = \mathrm{E}(\tau_X|\ T),$$

$$\mu_{YT} = \mathrm{E}(Y|\ T) = \mathrm{E}(\tau_Y|\ T), \qquad (2)$$

and

$$\mu_{AT} = \mathrm{E}(A|\ T) = \mathrm{E}(\tau_A|\ T).$$

A critical assumption of Levine's method is *congenericity*, which may be formulated as the two *population invariance assumptions*, LL1 and LL2, below in (3) and (4).

LL1: For any target population, *T,*

$$\tau_X = a\tau_A + b. \qquad (3)$$

LL2: For any target population, *T*,

$$\tau_Y = c\tau_A + d. \qquad (4)$$

In LL1 and LL2, the values of the linear parameters, *a*, *b*, *c,* and *d*, are assumed to be the same for any *T* of the synthetic form, so that the linear relations between the true scores of *X* and *Y* with *A* are *population invariant*. Assumptions LL1 and LL2 imply that for any *T*, the true scores of the three tests are perfectly correlated. This is the classical test theory way of asserting that the three tests measure the same thing but not necessarily in the same scale or with the same reliability.

The assumptions, LL1 and LL2, may be used to derive formulas for the means and standard deviations of *X* and *Y* on *T*. These then may be used to define the Levine linear-observed score equating function, $\mathrm{Lin}_{XY(L)}(x)$ in (6), below. The results are given in Kolen and Brennan (2004, p. 122) and make use of the reliability formulas derived by Angoff (1982).

Angoff (1982) derived useful estimates for the reliability ratios that make use of data that are available in the NEAT design. Angoff's (1982) estimates take different forms depending on whether *A* is internal or external to the two tests, *X* and *Y*.

In the rest of this paper, we assume that the Levine estimates, $\mu_{YT(L)}$, $\mu_{XT(L)}$, $\sigma_{XT(L)}$, and $\sigma_{YT(L)}$, of the means and standard deviations of *X* and *Y* on *T* are available.

In general, any linear equating function is formed from the first two moments of *X* and *Y* on *T* as

$$\mathrm{Lin}_{XY\,T}(x) = \mu_{YT} + (\sigma_{YT}/\sigma_{XT})(x - \mu_{XT}). \tag{5}$$

The Levine observed-score linear equating function is obtained from (5) when the first two moments of *X* and *Y* are estimated by the Levine estimates, as in (6) below:

$$\mathrm{Lin}_{XY\,T(L)}(x) = \mu_{YT(L)} + (\sigma_{YT(L)}/\sigma_{XT(L)})(x - \mu_{XT(L)}). \tag{6}$$

Even though it is restricted to be linear, the Levine linear function is often computed for comparison purposes with other linear methods. Under some circumstances it is more accurate than other linear equating methods (Petersen, Marco, & Stewart, 1982).

The Relation Between Linear and Equipercentile Equating Functions

Following von Davier et al. (2004a, 2004b), all observed-score equating functions linking *X* to *Y* on *T*, can be regarded as equipercentile equating functions that have the form shown in (7):

$$\mathrm{Equi}_{XY\,T}(x) = G_T^{-1}(F_T(x)), \tag{7}$$

where $F_T(x)$ and $G_T(y)$ are forms of the *cumulative distribution functions* (cdfs) of *X* and *Y* on *T*, and $y = G_T^{-1}(p)$ is the inverse function of $p = G_T(y)$. Different assumptions about $F_T(x)$ and $G_T(y)$ lead to different versions of $\mathrm{Equi}_{XY\,T}(x)$ and therefore to different observed-score equating functions.

Let μ_{XT}, μ_{YT}, σ_{XT}, and σ_{YT} denote the means and standard deviations of *X* and *Y* on *T* that are computed from $F_T(x)$ and $G_T(y)$, as in $\mu_{XT} = \int x dF_T(x)$, and so on. The linear equating function in (5) that uses the first two moments computed from $F_T(x)$ and $G_T(y)$ will be said to be compatible with $\mathrm{Equi}_{XY\,T}(x)$ in (7). It is the compatible version of $\mathrm{Lin}_{XY\,T}(x)$ that appears in

Theorem 1 below. We return to the issue of compatible linear and equipercentile equating functions in more detail later. Theorem 1 is proved in von Davier et al. (2004b) and connects the equipercentile function, $\mathrm{Equi}_{XY\,T}(x)$, in (7) to its compatible linear equating function, $\mathrm{Lin}_{XY\,T}(x)$, in (5).

Theorem 1: For any population, T, if $F_T(x)$ and $G_T(y)$ are continuous cdfs, and F_0 and G_0 are the standardized cdfs that determine the "shapes" of $F_T(x)$ and $G_T(y)$, that is, both F_0 and G_0 have mean 0 and variance 1 and

$$F_T(x) = F_0\left(\frac{x-\mu_{XT}}{\sigma_{XT}}\right) \text{ and } G_T(y) = G_0\left(\frac{y-\mu_{YT}}{\sigma_{YT}}\right), \tag{8}$$

then

$$\mathrm{Equi}_{XY\,T}(x) = G_T^{-1}(F_T(x)) = \mathrm{Lin}_{XY\,T}(x) + R(x), \tag{9}$$

where the remainder term, $R(x)$, *is equal to* $\sigma_{YT}\, r\left(\frac{x-\mu_{XT}}{\sigma_{XT}}\right)$, (10)

and r(z) is the function

$$r(z) = G_0^{-1}(F_0(z)) - z. \tag{11}$$

When $F_T(x)$ and $G_T(y)$ have the same shape, it follows that $r(z) = 0$ in (11) for all z, so that the remainder in (17) satisfies $R(x) = 0$, and thus, $\mathrm{Equi}_{XY\,T}(x) = \mathrm{Lin}_{XY\,T}(x)$.

Theorem 1 can be viewed as a sharpening of the well-known fact that when $F_T(x)$ and $G_T(y)$ have the same shape, the equipercentile equating function is identical to the linear equating function. It should be pointed out that the symmetry property of equating is preserved in Theorem 1.

It is important to recognize that, for the various methods used in the NEAT design, it is not always true that the means and standard deviations of X and Y used to compute $\mathrm{Lin}_{XY\,T}(x)$ are the same as those from $F_T(x)$ and $G_T(y)$ that are used in (7) to form $\mathrm{Equi}_{XY\,T}(x)$. The compatibility of a linear and equipercentile equating function depends on both the equating methods and how the continuization process for obtaining $F_T(x)$ and $G_T(y)$ is carried out.

The continuization method for KE/PSE insures that the means and standard deviations of $F_T(x)$ and $G_T(y)$ are the same as those of the underlying discrete distributions for any choice of bandwidth. In KE, $\text{Lin}_{XY\,T}(x)$ corresponds to large bandwidths, whereas $\text{Equi}_{XY\,T}(x)$ corresponds to smaller bandwidths that optimize a penalty function (von Davier et al., 2004b). Thus, in KE/PSE the four moments underlying $\text{Lin}_{XY\,T}(x)$ are the same as those of the $F_T(x)$ and $G_T(y)$ that underlie $\text{Equi}_{XY\,T}(x)$. Hence, for KE/PSE, the linear and equipercentile functions are compatible.

However, the traditional method of continuization by linear interpolation (Kolen & Brennan, 2004) does not reproduce both the mean and variance of the underlying discrete distribution. The piece-wise linear continuous cdf that the linear interpolation method produces is only guaranteed to reproduce the mean of the discrete distribution that underlies it. The variance of the continuized cdf is larger than that of the underlying discrete distribution by 1/12 (Holland & Thayer, 1989). Moreover, the four moments of *X* and *Y* on *T* that are implicitly used by the chain linear or the Tucker linear method are not necessarily the same, nor are they the same as those of the continuized cdfs of frequency estimation or the chain equipercentile methods. To our knowledge, there is, at best, an incomplete understanding of the compatibility of the various linear and equipercentile methods used in practice for the NEAT design.

The KE/PSE method has all the necessary ingredients for using the result of Theorem 1. Because of this, for KE/PSE we may calculate the function $r(z)$ in (11) directly without first forming F_0 and G_0. This computation is summarized in Theorem 2.

Theorem 2: If $\text{Equi}_{XY\,T}(x)$ and $\text{Lin}_{XY\,T}(x)$ in (5) and (7) are compatible, then $r(z)$ in (11) may be computed as

$$r(z) = \frac{1}{\sigma_{YT}}\{\text{Equi}_{XY\,T}(\mu_{XT} + \sigma_{XT}z) - \text{Lin}_{XY\,T}(\mu_{XT} + \sigma_{XT}z)\}. \qquad (12)$$

The proof of Theorem 2 simply solves for $r(z)$ using (9) and (10), so we omit it.

A General Proposal for Forming Hybrid Equipercentile Equating Functions

With this preparation, we are in a position to propose a way of obtaining a variety of hybrid equipercentile equating functions of the form (7) whose linear part is the linear Levine equating function in (6). The idea is to use (9) with the linear equating function being the Levine linear function, as shown in (13), below:

$$\mathrm{Lin}_{XY\,T}(x) = \mathrm{Lin}_{XY\,T(L)}(x) \qquad (13)$$

and the remainder function, $R(x)$, being computed from an $r(z)$ function found using (12) from some other appropriate equating method and the Levine estimates, $\mu_{XT(L)}$, $\sigma_{XT(L)}$, and $\sigma_{YT(L)}$.

Following this recipe, our proposed hybrid equipercentile Levine equating function has the form in (14):

$$\mathrm{Equi}_{XY\,T(L)}(x) = \mathrm{Lin}_{XY\,T(L)}(x) + \sigma_{YT(L)}\, r\left(\frac{x - \mu_{XT(L)}}{\sigma_{XT(L)}}\right). \qquad (14)$$

Equation (14) preserves the symmetry property that is required by equating functions (Dorans & Holland, 2000).

Using (12), we may express $\mathrm{Equi}_{XY\,T(L)}$ in terms of the Levine linear function, $\mathrm{Lin}_{XY\,T(L)}$, and the other two equating functions that were used as well. This is summarized in (15),

$$\mathrm{Equi}_{XY\,T(L)}(x) = \mathrm{Lin}_{\mathrm{XY}\,T(L)}(x) +$$

$$\frac{\sigma_{YT(L)}}{\sigma_{YT}}\left\{\mathrm{Equi}_{XY\,T}\left(\mu_{XT} + \frac{\sigma_{XT}}{\sigma_{XT(L)}}(x - \mu_{XT(L)}\right) - \mathrm{Lin}_{XY\,T}\left(\mu_{XT} + \frac{\sigma_{XT}}{\sigma_{XT(L)}}(x - \mu_{XT(L)})\right)\right\}. \qquad (15)$$

The argument of both $\mathrm{Lin}_{\boldsymbol{XY\,T}}$ and $\mathrm{Equi}_{\boldsymbol{XY\,T}}$ in (15),

$$\mu_{XT} + \frac{\sigma_{XT}}{\sigma_{XT(L)}}(x - \mu_{XT(L)}),$$

has the form of a linear equating function that links the Levine linear scale to that of the linear scale based on the moments, μ_{XT}, μ_{YT}, σ_{XT}, and σ_{YT}.

The Hybrid PSE-Levine Equipercentile Equating Function

In the KE version of PSE, the anchor test is used as a covariate on which the score probabilities for X and Y are poststratified and reweighted to obtain estimated score probabilities on T—$\{r_{jT}\}$ for X and $\{s_{kT}\}$ for Y. These are then continuized to produce two cdfs, $F_{T(PSE)}(x)$ and $G_{T(PSE)}(y)$. As mentioned earlier, because of the way KE continuization works, each of the two continuous cdfs has the same means and standard deviations as the corresponding discrete score probability distributions, $\{r_{jT}\}$ or $\{s_{kT}\}$. Thus, we can simply use $\{r_{jT}\}$ and $\{s_{kT}\}$ to obtain $\mu_{XT(PSE)}$, $\mu_{YT(PSE)}$, $\sigma_{XT(PSE)}$, and $\sigma_{YT(PSE)}$, via the usual definitions,

$$\mu_{XT(PSE)} = \sum_j x_j r_{jT} \text{ , } \mu_{YT(PSE)} = \sum_k y_k s_{kT} \text{ ,} \tag{16}$$

$$\sigma^2_{XT(PSE)} = \sum_j (x_j - \mu_{XT(PSE)})^2 r_{jT} \text{ , } \sigma^2_{YT(PSE)} = \sum_k (y_k - \mu_{YT(PSE)})^2 s_{kT} \text{ .} \tag{17}$$

Thus, for the KE version of PSE, forming integrals like $\int x dF_{XT(PSE)}(x)$ to compute $\mu_{XT(PSE)}$ and so on is unnecessary.

In order to use (15), it is necessary to have a way of calculating the KE/PSE functions, $\text{Equi}_{XY\,T(PSE)}(x)$ and $\text{Lin}_{XY\,T(PSE)}(x)$, for any value of x, not at just the scores values, $\{x_j\}$. We assume that this calculation is possible, though it may require modification of existing software. Then, values of $\mu_{XT(PSE)}$ and $\sigma_{XT(PSE)}$ are used as the values of μ_{XT}, σ_{XT} in (15) to compute the linear transformation

$$x^* = \mu_{XT(PSE)} + \frac{\sigma_{XT(PSE)}}{\sigma_{XT(L)}}(x - \mu_{XT(L)}). \tag{18}$$

In (18), x is a value at which we want to compute $\text{Equi}_{XY\,T(L)}(x)$ defined in (14) or (15). Finally, $\sigma_{YT(PSE)}$ is used as σ_{YT} to compute the nonlinear remainder term in (15) at the transformed value, x^*, as shown in (19),

$$\frac{\sigma_{YT(L)}}{\sigma_{YT(PSE)}}\{\text{Equi}_{XY\,T(PSE)}(x^*) - \text{Lin}_{XY\,T(PSE)}(x^*)\}, \tag{19}$$

and the result in (19) is then added to the Levine linear function, $\text{Lin}_{XY\,T(L)}(x)$, to compute $\text{Equi}_{XY\,T(L)}(x)$, as shown in (20),

$$\text{Equi}_{XY\,T(L)}(x) = \text{Lin}_{XY\,T(L)}(x) + \frac{\sigma_{YT(L)}}{\sigma_{YT(PSE)}}\{\text{Equi}_{XY\,T(PSE)}(x^*) - \text{Lin}_{XY\,T(PSE)}(x^*)\}. \tag{20}$$

The result in (20) is the PSE-Levine equipercentile equating function.

If the means and variances on T derived under the Levine assumptions are the same as the means and variances on T derived under the PSE assumptions, then (18) simplifies to the identity function, $x^* = x$, and equation (20) reduces to

$$\mathrm{Equi}_{XY\,T(L)}(x) = \mathrm{Lin}_{XY\,T(L)}(x) + \{\mathrm{Equi}_{XY\,T(PSE)}(x) - \mathrm{Lin}_{XY\,T(PSE)}(x)\}. \qquad (21)$$

It is an empirical question if such a simplification is realistic, but (21) only requires the computation of the difference between the two KE/PSE functions,

$$\mathrm{Equi}_{XY\,T(PSE)}(x) \text{ and } \mathrm{Lin}_{XY\,T(PSE)}(x).$$

Later in this paper we illustrate the ideas behind $\mathrm{Equi}_{XY\,T(L)}(x)$ using (21) as an approximate PSE-Levine equipercentile equating function.

A possibly more realistic approximation is to assume that the Levine and PSE variance estimates are identical, but not their estimates of the means. This leads to the following alternative approximation to the PSE-Levine equipercentile function:

$$\mathrm{Equi}_{XY\,T(L)}(x) = \mathrm{Lin}_{XY\,T(L)}(x) +$$

$$\{\mathrm{Equi}_{XY\,T(PSE)}(x + \delta) - \mathrm{Lin}_{XY\,T(PSE)}(x + \delta)\}, \qquad (22)$$

where, in (22) δ is the difference between the PSE and Levine estimates of the mean of X in T, $\delta = \mu_{XT(PSE)} - \mu_{XT(L)}$.

An Illustrative Example

Data

The data we use to illustrate our approach come from von Davier et al. (2006). Two unique 44-item pseudo-test scores, X and Y, and one 24-item, external-anchor test score, A, were carefully constructed from the item responses to a longer 120-item test. The 120-item test had been taken by two samples of examinees from two populations that differed in performance on this test. The mean total scores of the examinees taking the test at these two administrations, P and Q, differed by approximately one fourth of a standard deviation on the original 120-item test (see Table 2).

Table 2

Comparison of the Examinees at the Two Administrations on the Initial 120-Item Test

Administration	*P*	*Q*
Number of examinees	6,168	4,235
Mean	82.33	86.16
SD	16.04	14.19

The pseudo-tests, *X* and *Y*, were constructed in such a way that they were parallel in content but differed considerably in difficulty. On the combined group, the mean difference between *X* and *Y* was about 140% of the average standard deviation (see Table 3). When the test forms differ significantly in difficulty, the results from different equating methods also differ (von Davier et al., 2004a); hence, this design provides a good framework for investigating the differences in the equating methods and their assumptions. One might decide to use the term *linking* instead of *equating* in a practical situation, where the test forms exhibit massive differences in difficulty.

Table 3

Comparison of the Examinees at the Two Administrations on the Pseudo-Tests

Populations		*X*	*Y*
Examinees in *P*	Mean	36.4	28.0
(*n* = 4,237)	SD	4.8	6.3
Examinees in *Q*	Mean	35.1	26.6
(*n* = 6,168)	SD	5.7	6.7
Combined group, *T*	Mean	35.6	27.2
(*n* = 10,405)	SD	5.4	6.6

In addition, the anchor test was designed to be parallel in content but targeted at a difficulty level between *X* and *Y*. The reliabilities of *X* and *Y* were about 0.8; their correlations with the external anchor, *A*, were 0.78 in *P* and 0.76 in *Q*. The design table for the pseudo-test data is given in Table 4.

Table 4

The Design Table for the Pseudo-Test Data

	X	A	Y
P	✓	✓	✓
Q	✓	✓	✓

By ignoring the data for X in Q and Y in P, the scores from the pseudo-test data may be regarded as the NEAT design in Table 1, where the combined sample is regarded as from the synthetic population, $T = wP + (1 - w)\ Q$, with w proportional to the size of the sample from P. The data for X in Q and Y in P were used to augment this NEAT design to provide a criterion equating design that is not usually available. From Table 4 for the pseudo-test data, X and Y are seen to form a *single-group* (SG) *design* on T, the combined group. That is, every one in T has scores for both X and Y. This SG design provides a criterion equating that the NEAT design attempts to approximate. We used the full data set to estimate the KE SG design equipercentile function and treated it as the criterion equating for our analyses. Because this is not a simulation, “truth” is not known. Instead this paper uses a criterion equating that was constructed on the same population T as the equating functions of interest and through similar steps (presmoothing using loglinear models, continuization using linear interpolation) as the usual observed-score equating methods for the NEAT design. The equipercentile function was chosen because the two tests differ significantly in the shape of the distributions.

All of the equatings went from X to Y so that X plays the role of the new form and Y is the old form. The presmoothing of the data was accomplished by fitting appropriate loglinear models to the discrete score probability distributions (Holland & Thayer, 2000), as discussed in von Davier et al. (2006), who examined these data in detail.

The results of von Davier et al. (2006) indicated that an equipercentile version of the Levine observed-score equating function might be an appropriate equating function for these data. They found that the Levine linear function well approximated the SG linear criterion function. Table 5 shows the differences between the linear anchor equatings and the linear equating function in the SG design (considered the equating criterion for the linear equatings). More precisely, Table 5 shows (a) the maximum, minimum, and averages of these differences and (b) the root mean expected error (RMSE) of these differences. The RMSE, or error, is

$\mathrm{RMSE} = \sqrt{\overline{d}^2 + sd_d^2}$, where $\overline{d}$ is the mean of the differences of the equated scores ($d = a_i - b_i$, where a_i and b_i denote the equated scores of the score x_i by two different methods, respectively) and *sd* is the standard deviation of these differences.

Table 5

Summary Measures of Differences Between Linear KE/PSE, Tucker, Chain Linear, and Levine and the Criterion, SG Linear Equating

Summary	KE/PSE linear-criterion	Tucker-criterion	Chain linear-criterion	Levine-criterion
Mean difference	0.727	1.012	0.098	–0.152
SD difference	0.493	0.770	0.241	0.086
Max difference	1.564	2.302	0.501	–0.008
Min difference	–0.089	–0.279	–0.306	–0.296
RMSD difference	0.879	1.272	0.260	0.175

Note. KE = kernel equating, PSE = poststratification equating, RMSD = root mean squared deviation.

Figure 1 shows the difference between the Levine linear function and the SG linear (criterion) equating function. It indicates that the Levine function is a close approximation to the criterion linear equating based on the combined group.

However, because of the extreme difference in the difficulty of *X* and *Y*, a linear equating function is not satisfactory. This is seen in Figure 2, which displays the criterion equipercentile equating function that is decidedly not linear. Table 6 (von Davier et al., 2006) summarizes the differences from the nonlinear criterion of the anchor equating functions investigated in von Davier et al. (2006) at various score points.

Results

Due to software limitations at the time of this writing, we will illustrate only the approximate PSE-Levine equipercentile function that is given in (21) rather than the complete version of the PSE-Levine equipercentile function given by (18) and (19). Figure 3 graphs the three ingredients to (21)—(a) the Levine linear, (b) the KE/PSE equipercentile, and the (c) KE/PSE linear equating functions.

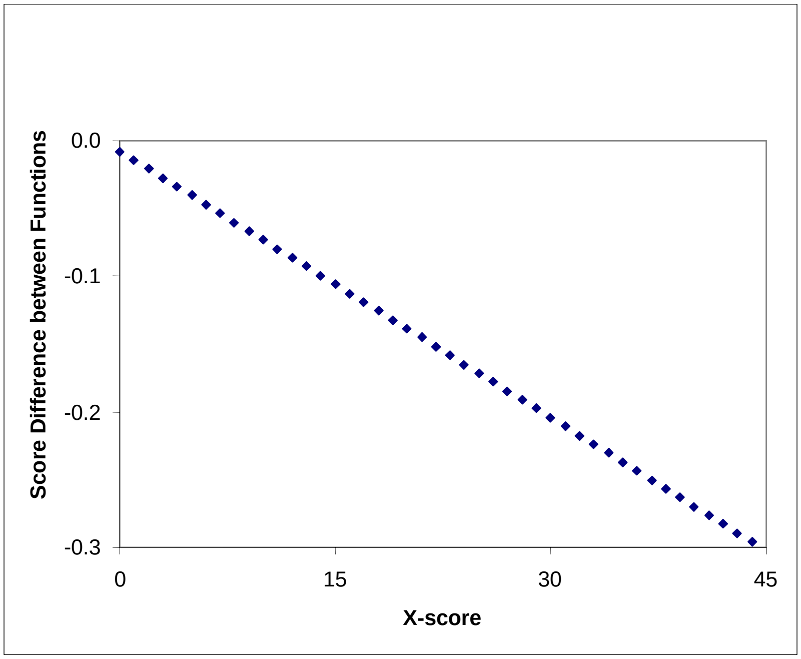

***Figure 1*. Graph of the difference between the Levine linear and the criterion single-group (SG) linear equating functions for the pseudo-test data.**

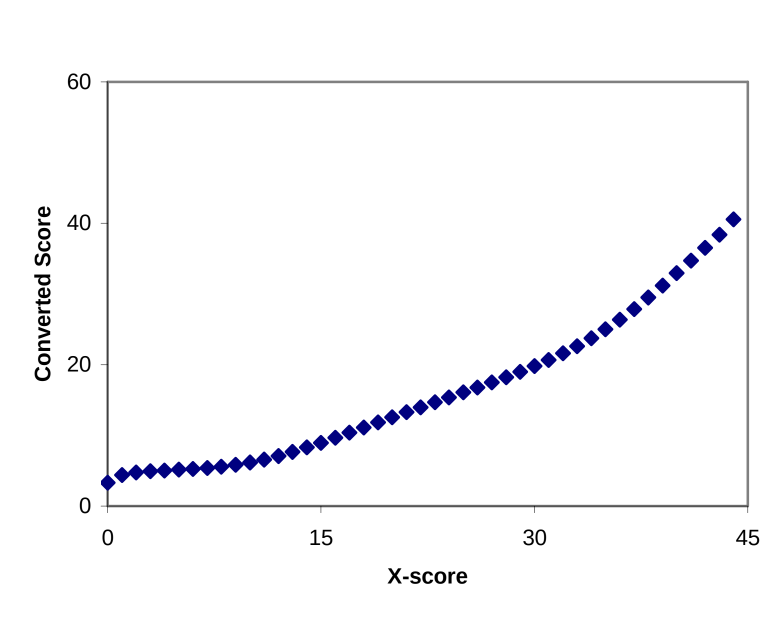

***Figure 2*. Graph of the criterion single-group (SG) kernel equating (KE) equipercentile equating function for the pseudo-test data.**

Table 6

Difference Between Each Anchor Equating and Criterion SG Nonlinear Equating for Selected Raw Scores

Difference from criterion equating	Raw score on Form *X*			
	25	30	35	40
Chain linear	–1.88	0.42	1.26	–0.68
KE/CE, large bandwidth	–1.57	0.61	1.33	–0.71
Tucker	–1.08	1.01	1.65	–0.49
KE/PSE, large bandwidth	–1.31	0.89	1.64	–0.39
Levine observed-score	–2.09	0.26	1.16	–0.71
Chain equipercentile	–0.17	–0.03	–0.06	–0.33
KE/CE, optimal bandwidth	–0.19	–0.03	–0.06	–0.32
Frequency estimation equipercentile	0.38	0.43	0.36	–0.07
KE/PSE, optimal bandwidth	0.35	0.44	0.37	–0.07

Note. KE = kernel equating, CE = chain equating, PSE = poststratification equating, SG = single group.

Again, it is clear from Figure 3 that the linear functions are not adequate to adjust for the extreme differences between the two pseudo- tests, *X* and *Y*, except possibly in the *X*-score range of about 20–40.

Figure 4 graphs both the approximate PSE-Levine equipercentile function using (21) and the criterion KE equipercentile SG equating function. They are remarkably close.

The difference between the two equating functions in Figure 4 is plotted in Figure 5. The largest differences are at the low end of the scale, where the approximate PSE-Levine equipercentile function underestimates the criterion equating function. The standard errors of the linear equating functions range from 0.53 (at Score 0) to 0.08 (from Scores 35–39) and show a similar pattern across the score range for all linear methods (see von Davier et al., 2006, p. 13).

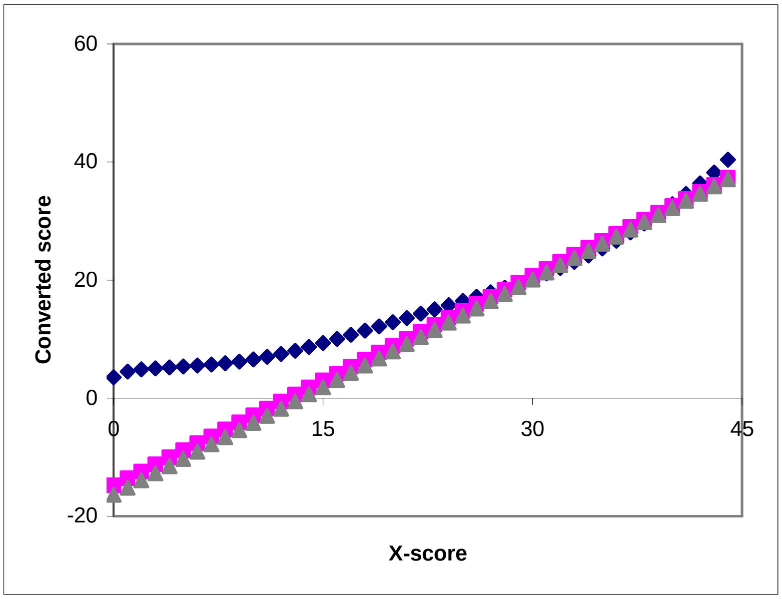

***Figure 3*. Graphs of the Levine linear, KE/PSE equipercentile, and KE/PSE linear equating functions.**

Note. KE = kernel equating; PSE = poststratification equating.

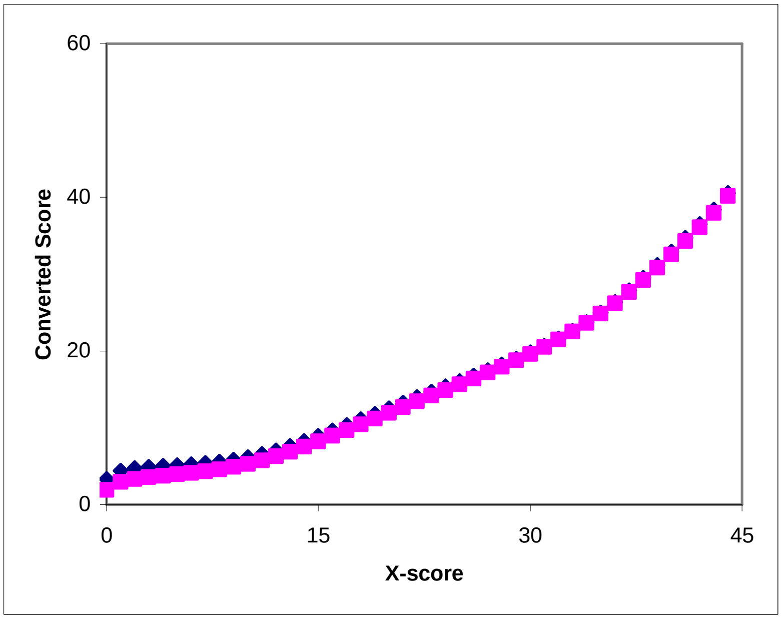

***Figure 4*. Graph of the approximate PSE-Levine equipercentile function from (28) and the criterion kernel equating (KE) equipercentile single-group (SG) function.**

Note. PSE = poststratification equating.

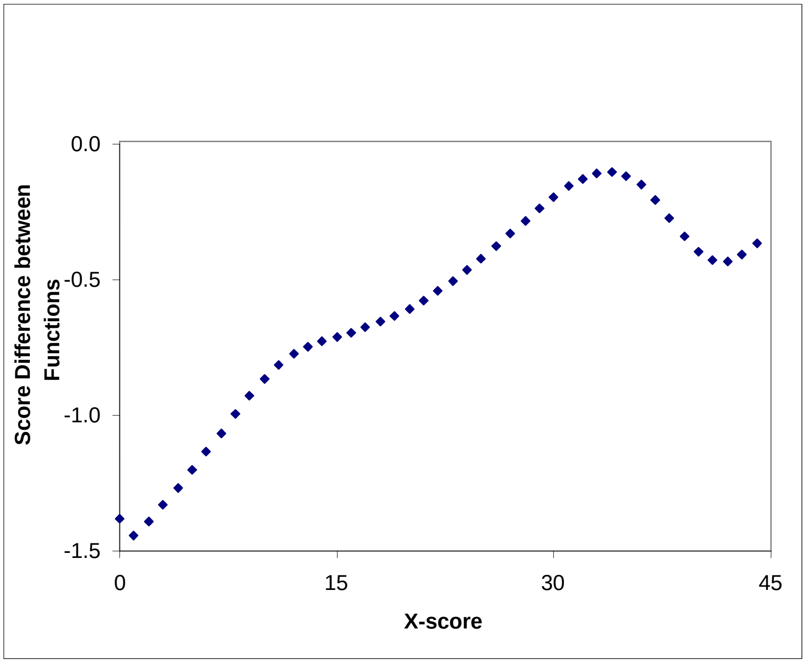

Figure 5**. Graph of the difference between the approximate PSE-Levine equipercentile equating function and the criterion equating function in Figure 4.**

Note. PSE = poststratification equating.

Discussion

We propose a general approach to creating a hybrid PSE-Levine equipercentile equating function that preserves the property of symmetry required of equating functions. The new function is based on a very basic decomposition of any equipercentile equating function into a linear and nonlinear part. We then suggest a hybrid that takes its linear part from the Levine linear function and its nonlinear part from some other equating method that includes compatible forms of equipercentile and linear functions. To the extent that the congeneric assumptions of the linear Levine function are satisfied and that the nonlinear part of the other equipercentile function is satisfactory, we would expect our proposal to be a useful addition to the methods for equating in the NEAT design.

We believe that the close agreement between the criterion equipercentile equating and the approximate version of the Levine equipercentile function found by using the KE/PSE equipercentile and linear functions suggests that it will be fruitful to pursue the approach indicated in this paper. Moreover, we think that the basic principle of KE, that the continuized cdfs should preserve at least the first two moments of the underlying discrete distribution, found a serious use in this application. Whereas it is the curvilinearity of equipercentile equating functions that usually gets the attention, the influence of the underlying means and variances should not be forgotten. These factors both locate and scale any equipercentile function and can have major effects on it.

Equation (15) allows for the possibility of a variety of different ways to combine the linear and nonlinear parts of different types of equating functions for the NEAT design. So far, we have explored only the combination of KE/PSE and the Levine linear method, but others are possible as well. For example, KE/CE may provide an alternative to KE/PSE in this regard. However, at this writing we are unclear whether the KE/CE equipercentile and KE/CE linear functions share the same underlying first two moments on a target population and are, therefore, compatible in the sense used here. This is a possible area for future research.

Our approach, especially (19), shows how important it is for equating software to allow for evaluating equating functions at values that are not just integer score values. Finally, we believe that investigations of the shapes of the $r(z)$ functions in (11) can be used to shed light on the differences between practical equipercentile equating methods. Computing and comparing the $r(z)$ functions for a variety of equipercentile methods appears to be a useful area for future research.

References

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–69). New York: Academic Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: ETS. (Originally published 1971 in *Educational measurement,* 2nd ed., pp. 508–600, by R. L. Thorndike, Ed., Washington, DC: American Council on Education).

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Research Rep. No. RR-89-07). Princeton, NJ: ETS.

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.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. RB-55-23). Princeton, NJ: ETS.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

Petersen, N. S., Marco, G. L, & Stewart, E. E. (1982). A test of adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and poststratification methods for observed-score equating: Their relationship to population invariance. In N. J. Dorans (Ed.), Assessing the population sensitivity of equating functions [Special issue]. *Journal of Educational Measurement, 41*(1), 15–32.

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

von Davier, A. A., Holland, P. W., Livingston, S. A.., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method. A special study with pseudotests constructed from real test data* (ETS Research Rep. No. RR-06-02). Princeton, NJ: ETS.