Listening. Learning. Leading.®

Research Report
ETS RR-13-38

The Kernel Levine Equipercentile Observed-Score Equating Function

Alina A. von Davier

Haiwen Chen

December 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

The Kernel Levine Equipercentile Observed-Score Equating Function

Alina A. von Davier and Haiwen Chen
Educational Testing Service, Princeton, New Jersey

December 2013

Action Editor: James Carlson

Reviewers: Sooyeon Kim and Rui Gao

Abstract

In the framework of the observed-score equating methods for the nonequivalent groups with anchor test design, there are 3 fundamentally different ways of using the information provided by the anchor scores to equate the scores of a new form to those of an old form. One method uses the anchor scores as a conditioning variable, such as the Tucker method and poststratification equating. A second way to use the anchor scores is as the middle link in a chain of linking relationships, such as chain linear equating and chain equating. The third way to use the anchor scores is in conjunction with the classical test theory, such as Levine observed-score equating and the newly created hybrid Levine equipercentile equating and poststratification equating based on true anchor scores. The purpose of this paper is to demonstrate that with real data, under certain conditions, hybrid Levine equipercentile equating and poststratification equating based on true anchor scores outperform both poststratification equating and chain equating.

Key words: nonequivalent groups with anchor test design, hybrid Levine equipercentile equating, poststratification equating based on true anchor scores, Levine observed-score equating

In the nonequivalent groups with anchor test (NEAT) design (also called the *common items* or *anchor test* 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), which 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 kernel Levine equipercentile under kernel equating (KE) framework was introduced in von Davier, Fournier-Zajac, and Holland (2007). Chen and Holland (2010); Chen, Livingston, and Holland (2011); Chen (2012); and Chen and Livingston (2012) used the same KE framework to develop a general version of Levine equating.

In her dissertation, Hou (2007) investigated one of the hybrid Levine equipercentile equatings (von Davier et al., 2007) and compared it to two classes of methods: chain equating and poststratification equating. She used simulated data generated by an item response theory (IRT) model with the conditions (80 cases) preset on combinations of five factors: sample size (two sizes), group proficiency difference (five cases), test length (two lengths), ratio of the number of common items to total test length (two ratios), and similarity of form difficulty (two cases). She concluded that hybrid Levine equipercentile equating yielded the smallest weighted absolute bias under almost all conditions (78 out of 80 times; Hou, 2007, p. 87). In particular, if the group proficiency difference is a combination of differences in the first two moments of the distributions, then the hybrid Levine equipercentile equating method performed best (p. 80). However, although very attractive theoretically, none of the kernel Levine equipercentile approaches seem to have been adopted in equating applications except for the equivalent equating method used in IRT equating (Chen, 2012).

The purpose of this paper is to summarize prior work on extensions of the Levine equating methods and argue for the practical benefits of these new methodologies. 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 (Mroch, Suh, Kane, & Ripkey, 2009; Petersen, Marco, & Stewart, 1982). In situations when a linear equating function is not satisfactory, an equipercentile version of the Levine function is desirable. There are several versions of equipercentile Levine equating. One is a *hybrid equating* function that combines linear and nonlinear equating functions in a systematic way that preserves the symmetry required of equating functions (von Davier et al., 2007); another is *poststratification*

equating on true anchor scores, with a relationship to Levine equating that is parallel to the relationship between poststratification equating and Tucker equating (Chen & Livingston, 2012). The general form of the Levine function will be soon available in KE Software at Educational Testing Service.

This paper discusses several ways to create an equipercentile version of the Levine linear observed-score equating method. It uses ideas from von Davier, Holland, and Thayer (2004b) and from Chen and Holland (2010) and exploits the general structure of the observed-score equating framework (von Davier, 2011, 2013). We present a general theoretical proposal and the results from two empirical studies. In one of the studies the results are derived under stronger assumptions than the general theory. The other study is an illustration of the methods with a real data set.

In NEAT design, the two test forms to be equated, *X* and *Y*, are taken, by two different samples of examinees; each sample is drawn from a different population, denoted here by *P* and *Q*. In this paper, *X* is called the *new* form and *Y* the *old* form, and the scores from *X* are placed on the scale of *Y*. In the NEAT design, it is not assumed that *P* and *Q* are similar in any way. To adjust for the ability differences in the two samples, a set of common items, *A*, is taken by the examinees from both samples. This data collection arrangement is shown in the design table (von Davier et al., 2004b), illustrated in Table 1.

Table 1

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

	X	A	Y
P	✓	✓	
Q		✓	✓

Note. Checkmarks denote that examinees in the samples indicated by the rows have scores on the test indicated by the columns.

If the scores of *A* are included in the scores of *X* (for Population *P*) or *Y* (for Population *Q*), then the anchor is called an internal anchor; otherwise, the anchor is called an external anchor.

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 scores, *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, Holland, and Thayer (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 that involve an implicit synthetic population, *T*, must hold.

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 in (5).

We will review Levine observed-score linear method in the next section. The following sections are: the hybrid equipercentile Levine equating; the poststratification equating on true

anchor scores (TAS) and its relations to both Levine observed-score equating (OSE) and IRT equating; comparisons of several equating methods on real data; and the discussion section.

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):

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

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 $0 \leq w \leq 1$. From (1), the basic equations in (2) follow for any *T* of this form:

$$\mu_{XT} = E(X \mid T) = E(\tau_X \mid T),$$
$$\mu_{YT} = E(Y \mid T) = E(\tau_Y \mid T), \tag{2}$$

and

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

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

LL1: For any target population, *T*,

$$\tau_X = a\tau_A + b. \tag{3}$$

LL2: For any target population, *T*,

$$\tau_Y = c\tau_A + d. \tag{4}$$

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, $\text{Lin}_{XY\,T(L)}(x)$ in (6). The results are given in Kolen and Brennan (2004, p. 122) and make use of the reliability formulas derived by Angoff (1982). Angoff derived useful estimates for the reliability ratios that make use of data that are available in the NEAT design. Angoff's 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, $\mu_{XT(L)}$, $\mu_{YT(L)}$, $\sigma_{XT(L)}$, 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

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

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).

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

Even though it is restricted to be linear, the Levine linear function is often computed for comparison purposes with other nonlinear methods. This is because under some circumstances it is more accurate than other linear equating methods (Mroch et al., 2009; Petersen et al., 1982).

Hybrid Equipercentile Levine Equating

In their paper, von Davier et al. (2007) proposed 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.

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):

$$\text{Equi}_{XYT}(x) = G_T^{-1}(F_T(x)), \tag{7}$$

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 $\text{Equi}_{XYT}(x)$ 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 $\text{Equi}_{XYT}(x)$ in (7). It is the compatible version of $\text{Lin}_{XYT}(x)$ 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, $\text{Equi}_{XYT}(x)$, in (7) to its compatible linear equating function, $\text{Lin}_{XYT}(x)$, in (5). This theorem has been known in other statistical applications as describing the shift model or location-scale model (Doksum & Sievers, 1976, p. 429).

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

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

where the remainder term, $R(x)$, is equal to

$$\sigma_{YT} r\left(\frac{x-\mu_{XT}}{\sigma_{XT}}\right), \tag{10}$$

and r(z) is the function

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

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 (9) satisfies $R(x) = 0$, and, thus, $\text{Equi}_{XY\,T}(x) = \text{Lin}_{XY\,T}(x)$.

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 $\text{Lin}_{XY\,T}(x)$ are the same as those from $F_T(x)$ and $G_T(y)$ that are used in (7) to form $\text{Equi}_{XY\,T}(x)$. The compatibility of a linear and an 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 $\text{Equi}_{XYT}(x)$ and $\text{Lin}_{XYT}(x)$ in (5) and (7) are compatible, then $r(z)$ in (11) may be computed as

$$r(z)=\frac{1}{\sigma_{YT}}\left[\text{Equi}_{XYT}(\mu_{XT}+\sigma_{XT}z)-\text{Lin}_{XYT}(\mu_{XT}+\sigma_{XT}z)\right]. \tag{12}$$

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:

$$\text{Lin}_{XYT}(x)=\text{Lin}_{XYT(L)}(x) \tag{13}$$

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):

$$\text{Equi}_{XYT(L)}(x)=\text{Lin}_{XYT(L)}(x)+\sigma_{YT(L)}r\left(\frac{x-\mu_{XT(L)}}{\sigma_{XT(L)}}\right). \tag{14}$$

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

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

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

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

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

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,

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

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

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

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

In (18), x is a value at which we want to compute $\text{Equi}_{XYT(L)}(x)$ 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),

$$\frac{\sigma_{YT(L)}}{\sigma_{YT(PSE)}}\left[\text{Equi}_{XYT(PSE)}(x^*) - \text{Lin}_{XYT(PSE)}(x^*)\right], \quad (19)$$

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

$$\text{Equi}_{XYT(L)}(x) = \text{Lin}_{XYT(L)}(x) + \frac{\sigma_{YT(L)}}{\sigma_{YT(PSE)}}\left[\text{Equi}_{XYT(PSE)}(x^*) - \text{Lin}_{XYT(PSE)}(x^*)\right]. \quad (20)$$

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 (20) reduces to

$$\text{Equi}_{XYT(L)}(x) = \text{Lin}_{XYT(L)}(x) + \left[\text{Equi}_{XYT(PSE)}(x) - \text{Lin}_{XYT(PSE)}(x)\right]. \quad (21)$$

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, $\text{Equi}_{XYT(PSE)}(x)$ and $\text{Lin}_{XYT(PSE)}(x)$.

More realistically, (21) stands if we assume that the difference between the two equipercentile functions is contained in the difference of their linear approximations.
Later in this paper, we illustrate the ideas behind $\text{Equi}_{XYT(L)}(x)$ using (21) as an approximate PSE-Levine equipercentile equating function.

Poststratification Equating Based on True Anchor Scores (PSE-TAS)

In Chen and Livingston (2012), a different equipercentile Levine equating was constructed, using a generalized version of the kernel equating framework introduced in von Davier (2011, 2013).

Observed-Score Equating (OSE) Framework

The OSE framework was derived from the KE framework, and it was presented in von Davier (2011, 2013). The OSE framework has five steps, and it also includes Theorem 1: Presmoothing. Presmoothing can be done by using loglinear smoothing (the default choice in the KE framework), IRT models (a discussion will be given later), spline functions, or other models. Some models, such as IRT models, can produce presmoothed distributions of variables of either observed scores or true scores, or both. Some considerations include the following:

- *Estimating the score probabilities on the target population.* Here, a specified method will be used. For NEAT designs, the common method is either CE or PSE. The results are either two discrete univariate distributions for PSE or four distributions for CE. Local equating can be also employed (Wiberg, van der Linden, & von Davier, in press).
- *Continuization.* A Gaussian kernel is often used to transfer the discrete distributions into continuous ones. The choice of a parameter, called bandwidth, will determine whether the equating is linear (large bandwidth) or curvilinear (small bandwidth) in the following step. However, other kernel choices or other continuization methods are possible (see von Davier, 2011, for details).
- *Computing the equating function from the equipercentile equating on the continuized distributions.*
- *Computing the standard error of equating and related accuracy measures.* More and more evidence suggests that distributions of other types beside observed scores may be relevant also to a given equating task, and the framework can be applied to them as well (see the appendix).

The OSE framework unifies the whole equating process, where details can be studied more closely to reveal what really makes two equating methods different. One can see that if an equating process can be put under the framework, it can only be different from another equating

within the same framework in three areas: model for fitting the data (in Step 1), basic equating procedure (in Step 2), and choice of the continuization (in Step 3).

The construction of PSE-TAS. First, let us recall the definition of PSE. Let $f(X|a_j)$ be the conditional distribution of scores in Test X for examinees having anchor score a_j, $\{p(a_j)\}$ and $\{q(a_j)\}$ be the anchor score distributions of examinees taking Test X and Test Y, respectively, then the score distribution of X on synthetic population T is:

$$f_T(X) = \sum_j f(X \mid a_j)\left[wp(a_j) + (1-w)q(a_j) \right]. \quad (22)$$

Hence, $F_T(X)$, the cdf of X on T, can be computed accordingly. Similarly, we can get $G_T(Y)$, the cdf of Y on $\boldsymbol{T}$. Then PSE from X to Y on $\boldsymbol{T}$ is the equipercentile equating from $F_T(X)$ to $G_T(X)$. If we replace the anchor test A by its true score component τ_A in this equating construction, we get a new score distribution of X on $\boldsymbol{T}$:

$$f_{T,\tau_A}(X) = \sum_j f(X \mid \tau_{a_j})\left[wp(\tau_{a_j}) + (1-w)q(\tau_{a_j}) \right]. \quad (23)$$

where τ_{a_j} is a value of the true anchor score, and a new cdf of X on T, $F_{T,\tau_A}(X)$. Similarly, $G_{T,\tau_A}(Y)$ can be constructed, and the poststratification equating based on true anchor scores (PSE-TAS) is the equipercentile equating from $F_{T,\tau_A}(X)$ to $G_{T,\tau_A}(Y)$, which is the modified PSE (Wang & Brennan, 2007) conceptually.

The following theorem is proved in Chen and Livingston (2012):

Theorem 3: If all following conditions are satisfied:

In Population P, the conditional mean of X on true anchor τ_A is a linear function of τ_a, and the conditional covariance of X is constant on τ_A,

both the conditional mean of Y and the conditional covariance of Y on τ_A have the same properties in Population Q,

both τ_X and τ_Y are correlated perfectly with τ_A, and

the variance of the error component of A is population invariant,

then the linear equating from $F_{T,\tau_A}(X)$ to $G_{T,\tau_A}(Y)$ is the Levine observed-score equating.

Remark: Conditions 1–4 are equivalent to the assumptions for Levine observed-score equating. Because both of the score distributions, $f_{T,\tau_A}(X)$ and $g_{T,\tau_A}(Y)$, are defined explicitly, the assumptions for the Levine equating are translated as the conditions that can be verified.

This method was (re)discovered (the first version appeared in Wang and Brennan, 2007) when the first author studied the relationship between IRT observed-score equating and Levine observed-score equating, using the kernel equating framework. By presmoothing data with IRT models, and making both equatings linear, the simulation study shows that these two methods produce identical results (Chen, 2012). Chen concluded that IRT observed-score equating is the poststratification equating based on true anchor scores on data presmoothed by IRT models.

There is a technical issue with PSE-TAS: No general method is available to get the bivariate distributions whose marginal distribution on the main test is of the raw scores, but whose marginal distribution on the anchor test is of the true scores in the classical test theory model. The alternative is to use a linear transformation on the bivariate distribution that preserves the means of both marginal distributions but changes standard deviations with calculated ratios. One such method is called PSEκ (Chen & Holland, 2010; Chen et al. 2011), where κ is a number in [0, 1]. PSE_0 is the PSE method, while PSE_1 is equivalent to PSE-TAS but may have discrepancies on the points near both ends of the score range. PSEκ can be applied to test forms either with an internal anchor or an external anchor. Another method is called modified PSE (Wang & Brennan, 2007). Their study only applied the method to test forms with internal anchors, and the computation of the ratios to change the standard deviations is also different.

Comparisons of Levine Methods With CE and PSE

Extensive research has been done to compare the classical equating methods and IRT equating methods. Chen (2012) provides an extensive but still incomplete list of research papers in this area. However, many papers in the list do not answer the question of which method is closer to the true equating. Some papers use IRT models for simulated data, and consequently use IRT equating as the *true* equating. As pointed out in Chen (2012), IRT equating can be regarded as a curvilinear Levine equating with data presmoothed with an IRT model. Using it as a criterion will be definitely in favor of the Levine-type methods discussed in this paper. Therefore, we want to use real data and show that**,** under certain circumstances, Levine-type

methods will do better than other commonly used methods, although**,** in many cases, the opposite is true.

Why do different equating methods produce different equating results, particularly in terms of equated means? Research results indicate that many factors (content format, content difficulty, test form length, population ability, etc.) will contribute to the differences of the methods. Chen et al. (2011) make several assumptions to eliminate the impact of factors other than the population ability difference and find that Levine observed-score equating has the highest equated scores for *X* if the mean anchor test scores on **P**opulation *P* is higher than that on Population *Q*. Tucker equating resulted in the lowest equated scores**,** and the chain linear method produced scores between the other two methods. This phenomenon is well known to psychometricians and is mentioned in several research papers (e.g., see Holland, Sinharay, von Davier, & Han**,** 2008). Therefore, to demonstrate that Levine methods may work better than other methods, we need to construct two groups for which:

- There is a notable difference on the means of their anchor test scores (the bigger the difference**,** the better the performance of the Levine methods).
- The score distributions satisfy the assumptions for the traditional Levine observed-score equating.

In this paper, we will consider two circumstances, both represented by operational data. The first data set contains manipulated data as explained below. The second data set is a real operational data set.

Data Analysis

Design of the Comparisons

One way to construct a NEAT design such that the equatings on the design can be checked against a criterion is to split a long test with a large sample of examinees into two populations (*P* and *Q*) and three tests (two pseudo tests and an anchor test) as shown in Table 2, where the details of selections of *P*, *Q*, *X*, *Y*, and *A* will be discussed later.

Table 2

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*. There is a second NEAT design that ignores the data for *X* in *P* and *Y* in *Q*. Then *w* is proportional to the size of *Q*. 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 2, for the pseudo-test data, *X* and *Y* are seen to form a *single-group* (SG) *design* on *T*, the combined group. That is, everyone 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 Gaussian kernel) 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.

Data Set 1

The data we use to illustrate our approach come from von Davier et al. (2006). The 120-item test had been taken by more than 10,000 examinees.

First, Populations *P* and *Q* were constructed in such way that their performances are very different on the test and the conditional distributions based on their abilities are equivalent.

An IRT model is fitted with all test scores to determine each examinee's ability (θ). The ability distribution is divided into 41 intervals centered at points from -4 to 4 with an increment of 0.2. Formula θ/8 + ½ is used to determine that how many examinees in each ability band are in Population *Q*. For example, if there are 200 examinees in the band that θ = 1, then 200*(1/8+1/2) = 125 examinees in the band are randomly assigned to Population *Q*. The statistics of *P* and *Q* are given in Table 3.

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 form a longer 120-item test. 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). One might decide to use the term *linking* rather than *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

Population	Statistic	X	Y	A
Examinees in *P*	Mean	34.3	25.5	15.5
(*n* = 5,187)	SD	5.7	6.5	4.2
Examinees in *Q*	Mean	36.9	28.8	17.4
(*n* = 5,213)	SD	4.6	6.1	3.8
Combined group, *T*	Mean	35.6	27.2	16.4
(*n* = 10,400)	SD	5.4	6.6	4.1

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 0.78; their correlations with the external anchor, *A*, were from 0.74 to 0.77, on Populations *P* and *Q*, respectively.

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. This is due to the extreme difference in the difficulty of *X* and *Y*. One can see in Figure 1 that the criterion equipercentile equating function is decidedly not linear.

The ranges of scores for *X* and *Y* were also modified to exclude many almost-empty cells because the equatings are very unstable in such ranges. The modified *X* score range is [13, 44] and the modified *Y* score range is [8, 44]. Only five records out of 10,405 were taken out of the original distribution.

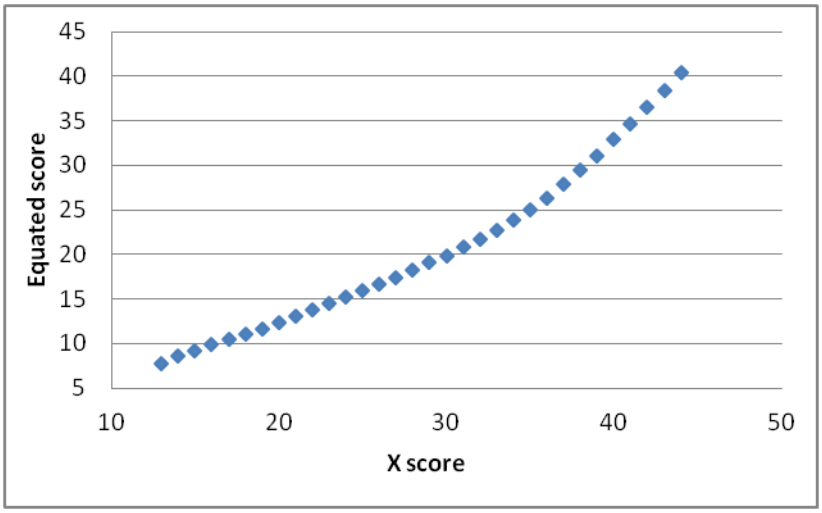

Figure 1. **The criterion single-group (SG) kernel equating (KE) equipercentile equating function for the pseudo-test data.**

Results

Four equating methods were compared against SG equating. They were KE/PSE, KE/CE, ***h***ybrid Levine (Equation 21), and PSE_1 (Chen & Holland, 2010). To distinguish the impacts of group ability difference from the test form difficulty difference, all four equating methods were compared both from X_P (Population P taking Test X) to Y_Q and from X_Q to Y_P.

The loglinear model for all bivariate distributions is (5, 5, 1). Since *P* and *Q* are of similar size, *w* is 0.5 for equating either from X_P to Y_Q or from X_Q to Y_P.

Table 4 shows (a) the maximums, minimums, averages, and standard deviations of these differences and (b) the root mean squared errors (RMSE) of these differences. The RMSE, or error, is defined as $\sqrt{\overline{d}^2 + sd_d^2}$, where $\overline{d}$ is the mean of the differences of the equated scores ($d_i = a_i - b_i$, 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. All means, standard deviations, and RMSEs of the differences were calculated on uniform distributions.

Table 4

Summary Measures of Differences Between KE/PSE, KE/CE, Hybrid Levine, and PSE_1 and the Criterion, SG Linear Equating, Both From X_P to Y_Q and X_Q to Y_P

Summary	KE/PSE criterion		KE/CE criterion		Hybrid Levine criterion		PSE_1 criterion	
NEAT type	*P* to *Q*	*Q* to *P*	*P* to *Q*	*Q* to *P*	*P* to *Q*	*Q* to *P*	*P* to *Q*	*Q* to *P*
Mean difference	1.07	-1.01	0.47	-0.46	0.07	-0.38	0.36	–0.24
SD difference	0.19	0.17	0.27	0.25	0.18	0.24	0.49	0.50
Max difference	1.32	-0.47	0.92	-0.20	0.29	0.05	1.47	0.89
Min difference	0.53	-1.29	0.00	-0.95	-0.52	–0.82	-1.11	–1.60
RMSE difference	1.09	1.03	0.54	0.53	0.19	0.45	0.61	0.55

Note. CE = chain equating, KE = kernel equating, NEAT = nonequivalent groups with anchor test, PSE = poststratification equating, RMSE = root mean squared error.

Figure 2 shows the differences between the four NEAT equating functions and the SG (criterion) equating function from X_P to Y_Q. It indicates that both Levine functions are close approximations to the criterion equating based on the combined group, although PSE_1 exhibits the undesirable trend at both ends of the range that was mentioned before. Since the mean of A_P is smaller than A_Q, one can see that the Levine type equatings get the lowest equated scores, KE/CE has the middle equated scores, and KE/PSE has the highest equated scores—and the curves are almost parallel. This phenomenon has been seen by many psychometricians and has been discussed in Chen et al. (2011) for their linear counterparts: Levine observed-score equating, chain linear, and Tucker equating.

Figure 3 shows the differences between the four NEAT equating functions and the SG (criterion) equating function from X_Q to Y_P. This time, the trend is reversed, since the mean of A_Q is larger than A_P. Overall, the hybrid Levine method still outperforms other equating methods, but for the central range of X, PSE_1 has the best result.

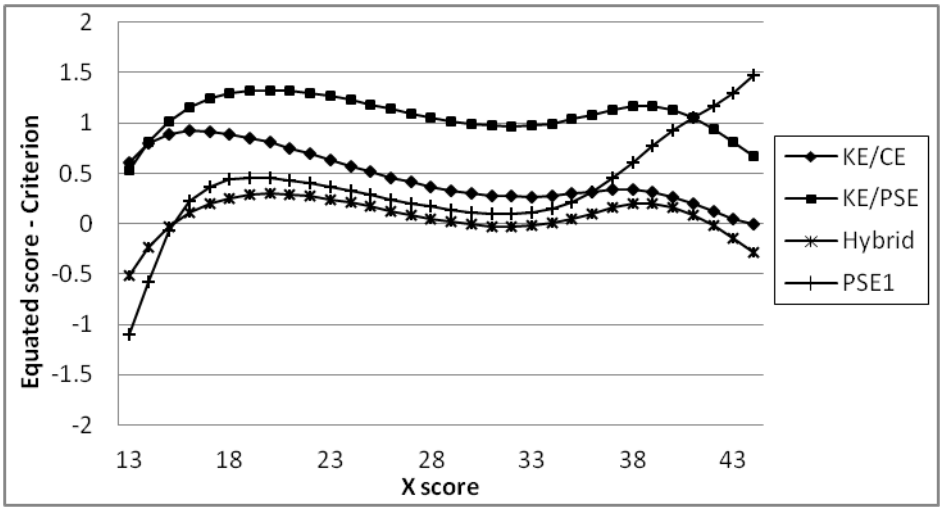

***Figure 2*. Differences between the four four nonequivalent groups with anchor test (NEAT) equatings from X_P to Y_Q and the criterion single-group (SG) equating functions for the pseudo-test data. CE = chain equating. KE = kernel equating, PSE = poststratification equating.**

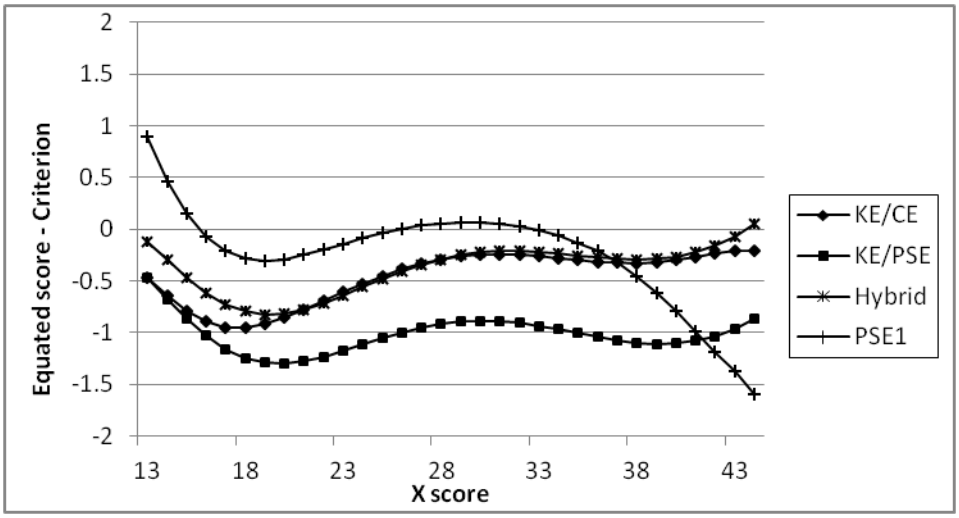

***Figure 3*. Differences between the four nonequivalent groups with anchor test (NEAT) equatings from X_Q to Y_P and the criterion single-group (SG) equating functions for the pseudo-test data. CE = chain equating, KE = kernel equating, PSE = poststratification equating.**

The approximate PSE-Levine equipercentile function using (21) and the criterion KE equipercentile SG equating function are remarkably close, followed by PSE_1, then KE/CE and finally KE/PSE. But it should be mentioned again that the data for this study are constructed to satisfy the assumptions for Levine observed-score equating. If we construct the data differently, other methods may prevail.

Data Set 2

Finally, using real data from a teacher licensing test, we will demonstrate that if the two groups in the NEAT design are not too far apart, the Levine methods will produce similar results to other equating methods.

The test has 91 items with 29 external anchor items. The old form was given in 2010, while the new form was given in 2011. Table 5 gives the statistics of both forms.

Table 5

Statistics of the New and Old Forms in a Teacher Licensing Test

Form	No. of examinees	Test mean	Test SD	Anchor Mean	Anchor SD
New	1,258	65.0	11.5	19.5	4.7
Old	4,948	65.3	11.7	20.0	4.6

Note. The difference between two anchor means is only a 0.5 raw score point.

Results

Four equating methods were applied to the data. The results are shown in Figure 4.

The score range is restricted to [15, 91] to avoid unstable equated values, since no test taker received a score less than 25.

The anchor mean for the new form is lower than for the old form. Since PSE, CE, and both hybrid Levine and PSE_1 are the curvilinear forms of Tucker equating, chain linear, and Levine observed-score equating, respectively, following the argument in Chen et al. (2011), it is not surprising to see that KE/PSE has the highest equated scores in general, followed by KE/CE, and then by hybrid Levine and PSE_1, although they are much closer in this case than in the previous example.

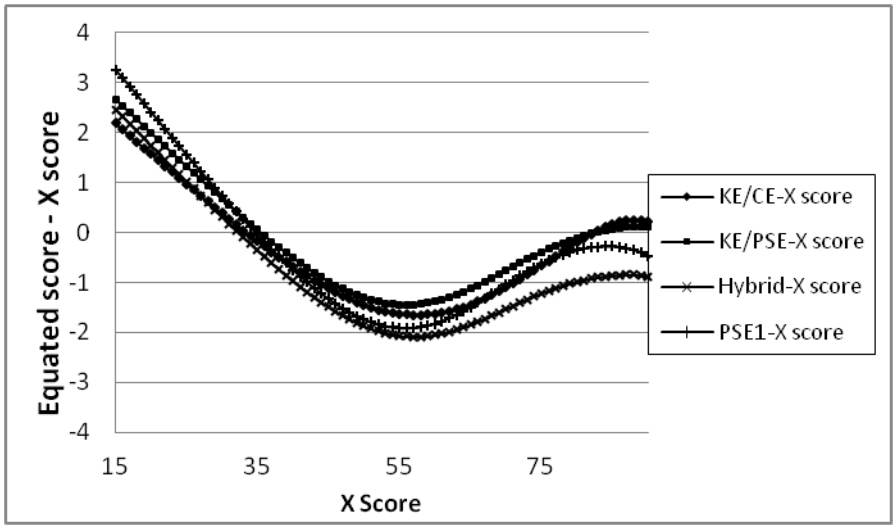

***Figure 4*. Graph of the differences between the four nonequivalent groups with anchor test (NEAT) equatings and the X scores. CE = chain equating, KE = kernel equating, PSE = poststratification equating.**

Discussion

Von Davier et al. (2007) proposed 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. While 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 unsure 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. 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.

Starting from Chen and Holland (2009), under the KE framework, several Levine equating methods have been created. These methods have natural relations with their linear counterparts. In particular, the relationship between PSE-TAS and Levine-OSE is almost identical to the relationship between PSE and Tucker equating (see Braun & Holland, 1982, for the second relationship). The only difference is that the first pair only use true anchor scores in their formulations while the second pair only use observed anchor scores.

The work in Chen et al. (2011) connects the dots among all established equating methods for NEAT designs. The three most well-known (linear) equating methods—Levine, chain, and Tucker—are ordered in favor of the higher ability group. Along with their nonlinear counterparts, a family of equating methods is created in Chen (2011). Many new equating methods are created while the older ones have found their equivalences in the family. More work is planned in this direction.

A technical issue has to be resolved for PSE-TAS and the majority of the methods in Chen (2011) to be useful in practice. Although PSEκ has a good approximation to each member in the family (Chen, 2011) with a specified κ, the discrepancies at the end points (shown in Figures 2 and 3) make it less desirable than another method, for example, hybrid Levine.

Before the introduction of the OSE framework (von Davier, 2011, 2013), the equating methods for NEAT designs appeared disconnected. The framework establishes connections between linear equatings and equipercentile equatings. Most importantly, it builds a system to classify equatings by dividing the whole process into steps where, at each step, the specified properties of an equating can be studied in great detail. With the generalized framework, many new equatings for NEAT designs can also be covered in the system, such as local equating (Wiberg et al., in press). Moreover, the distributions used can be of the observed scores, of the true scores, or one of the true scores and one of the observed scores. When two equatings are compared, the differences are displayed within each specification. For example, using IRT-OSE as the benchmark to compare CE with PSE is a biased comparison against PSE (Chen et al., 2011). A classification table with the specifications on some equating methods mentioned in this paper is provided in the appendix.

Future research should address several key issues. The first and most important one is how to determine the best equating method in practice. The newly created PSE-TAS gives researchers an insight in the equating process and offers a different kind of equating criterion for choosing the right equating method. The second issue is how to develop models/procedures to fit the data with true (anchor) scores well, particularly at the end score regions. The applications of the new data models are numerous. Not only are they needed for any equating methods that require true score distributions, they can also be used for direct computations of any statistics associated with the true scores.

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, NY: Academic Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

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, NY: Academic Press.

Chen, H. (2011, April). *A generalized linear equating and a generalized post-stratification equating both based on partial error anchor scores and their relationship*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Chen, H. (2012). A comparison between linear IRT observed score equating and Levine observed score equating under the generalized kernel equating framework. *Journal of Educational Measurement, 49,* 269–284.

Chen, H., & Holland, P. (2009). *Construction of chained true score equipercentile equatings under the KE framework and their relationship to Levine true score equating* (Research Report No. RR-09-24). Princeton, NJ: Educational Testing Service.

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika, 75,* 542–557.

Chen, H., & Livingston, S. A. (2012). *Post-stratification equating based on true anchor scores and its relationship to Levine observed score equating*. Manuscript submitted for publication.

Chen, H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 185–200). New York, NY: Springer-Verlag.

Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika, 63,* 421–434.

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

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45,* 17–43.

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

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.

Hou, J. (2007). *Effectiveness of the hybrid Levine equipercentile and modified frequency estimation equating methods under the common-item nonequivalent groups design* (Unpublished doctoral dissertation). University of Iowa, Iowa City.

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

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

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

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement, 7,* 174–193.

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, NY: Academic Press.

von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier, (Ed.), *Statistical models for test equating, scaling and linking* (pp.1–17). New York, NY: Springer-Verlag.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*(4), 605–623.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating* (Research Report No. RR-07-14). Princeton, NJ: Educational Testing Service.

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* (Research Report No. RR-06-02). Princeton, NJ: Educational Testing Service.

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,* 15–32.

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

Wang, T., & Brennan, R. L. (2007, April). *A modified frequency estimation equating method for the common-item non-equivalent groups design*. Paper presented at the meeting of the National Council of Measurement in Education, Chicago, IL.

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (in press). Local kernel observed-score equating. *Journal of Educational Measurement.*

Appendix

Classification Table of Some Equating Methods

Name	Presmoothing model	Score type	Design	Linearity
Linear IRT-OSE	IRT	(X, T_V), (Y, T_V)	PSE	Linear
IRT-OSE	IRT	(X, T_V), (Y, T_V)	PSE	Nonlinear
Linear IRT-TSE	IRT	(T_X, T_V), (T_Y, T_V)	PSE/CE	Linear
IRT-TSE	IRT	(T_X, T_V), (T_Y, T_V)	PSE/CE	Nonlinear
(IRT-Tucker)	IRT	(X, V), (Y, V)	PSE	Linear
(IRT-PSE)	IRT	(X, V), (Y, V)	PSE	Nonlinear
(IRT-chain linear)	IRT	(X, V), (Y, V)	CE	Linear
(IRT-CE)	IRT	(X, V), (Y, V)	CE	Nonlinear
Tucker equating [a]	Not specified	(X, V), (Y, V)	PSE	Linear
PSE	Not specified	(X, V), (Y, V)	PSE	Nonlinear
Chain linear	Not specified	(X, V), (Y, V)	CE	Linear
CE	Not specified	(X, V), (Y, V)	CE	Nonlinear
Levine observed-score equating [a]	Not specified	(X, T_V), (Y, T_V)	PSE	Linear
(PSE-TAS) [b]	Not specified	(X, T_V), (Y, T_V)	PSE	Nonlinear
Levine-TSE	Not specified	(T_X, T_V), (T_Y, T_V)	PSE/CE	Linear
(Chain TSE equating)[c]	Not specified	(T_X, T_V), (T_Y, T_V)	PSE/CE	Nonlinear

Note. Names in parentheses are suggested. CE = chain equating, IRT = item response theory, OSE = observed-score equating, PSE = poststratification equating, TAS = true anchor scores, TSE = true score equating.

[a] Both the Tucker method and Levine observed-score equating are approximated within the specified processes. [b] See Chen and Livingston (2012). [c] See Chen and Holland (2009).