Chapter 12
Generalized Equating Functions for NEAT Designs

Haiwen H. Chen, Samuel A. Livingston, and Paul W. Holland

12.1 Introduction

The purpose of this chapter is to introduce generalized equating functions for the equating of test scores through an anchor. Depending on the choice of parameter values, the generalized equating function can perform either linear equating or equipercentile equating, either by poststratification on the anchor or by chained linking through the anchor.

The generalized equating functions can represent either linear or equipercentile equating because they are based on the kernel equating procedure (von Davier, Holland, & Thayer, 2004b), which allows the user to choose a bandwidth for converting the discrete distributions of scores into continuous distributions. A large bandwidth causes the equating to be linear; a small bandwidth results in equipercentile equating. In addition, the generalized equating functions translates the choice of assumptions about the test and the anchor—the choice that leads to one or another of the familiar anchor equating methods—into the choice of a value for a new single parameter, called κ. For example, with a large bandwidth, one value of the κ parameter will produce Tucker equating; another value will produce chained linear equating; still another will produce Levine equating (Levine, 1955). Those same three values for κ, used with a small bandwidth, will produce frequency estimation equipercentile equating, chained equipercentile equating, and a nonlinear method analogous to Levine equating. The κ parameter also can take on infinitely many other values, producing a family of equating methods that includes the methods mentioned above as special cases.

H.H. Chen (✉) and S.A. Livingston,
Educational Testing Service, Rosedale Rd, Princeton, New Jersey 08541, USA
e-mail: hchen@ets.org

P.W. Holland
Paul Holland Consulting Corporation, 200 4th Ave South, Apt 100, St Petersburg FL 33701, USA
e-mail: pholland@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_12,

Section 12.2 of this chapter presents some of the terminology that will be used. Section 12.3 presents a uniform way to reproduce the three common linear equating methods: (a) the chained linear method, (b) the Tucker method, and (c) the Levine method. Section 12.4 presents the generalized equating function by poststratification on the anchor and shows how it can produce the familiar linear and equipercentile equating methods. Section 12.5 provides an example of the application of the generalized equating function to simulated data derived from an actual data set, varying the value of the κ parameter.

12.2 Terminology

The equating of scores on two different forms of a test is often accomplished on the basis of data collected when the groups of examinees taking the two forms are not of equal ability but are linked by a common "anchor" test taken by both groups. This data collection plan is often referred to as the nonequivalent groups with anchor test (NEAT) design. In this chapter, X and Y will refer to the scores on the two test forms; A will refer to the score on the anchor test. The examinees taking the two forms will be assumed to be sampled from different populations, referred to as P and Q (corresponding to test forms X and Y, respectively).

Several methods have been proposed for equating test scores on the basis of data from a NEAT design. Some of those methods constrain the equating relationship to be of the form $Y=\alpha+\beta X$. Those methods will be referred to as *linear equating* methods. Other methods do not impose this constraint; instead, they estimate the function that transforms the distribution of X into the distribution of Y in some specified population of examinees. Those methods will be referred to as *nonlinear equating* methods. The specified population will be referred to as $\boldsymbol{S}$ and is assumed to be a composite of populations P and Q, represented in the ratio w to $(1 - w)$.

To equate test scores on the basis of data from a NEAT design, it is necessary to assume that some characteristics of the bivariate distributions of test and anchor scores are population invariant—that they are the same in populations P, Q, and S. One common assumption is that the conditional distributions of X and Y, given A, are population invariant. That assumption makes it possible to estimate the distributions of scores X and Y in population S and then use those estimated distributions to equate X to Y. Equating on the basis of this assumption will be referred to as *poststratification equating*. The linear version of poststratification equating is the Braun-Holland method (Braun & Holland, 1982). Other linear equating methods based on similar assumptions include the Tucker method and the Levine method (Kolen & Brennan, 2004, pp. 105–132). The nonlinear version of poststratification equating (described by Angoff, 1971/1984) is commonly known as "frequency estimation" (p. 113) equating.

An alternative set of assumptions is that the symmetric linking relationships of X to A and of Y to A are population invariant. Equating methods based on these assumptions link score X to score A by assuming the linking relationship in population S to

be the same as in population P; they then link score A to score Y by assuming the linking relationship in population S to be the same as in population Q. These methods will be referred to as *chained equating*. The linear and nonlinear versions of chained equating are commonly known as *chained linear* and *chained equipercentile equating*.

12.3 A General Form for Tucker, Levine, and Chained Linear Equating

The basic equation for the linear equating of score X to score Y by estimating the means and standard deviations of X and Y in population $\boldsymbol{S}$ is Equation 12.1:

$$y = f(x) = \frac{\sigma_S(Y)}{\sigma_S(X)}[x - \mu_S(X)] + \mu_S(Y). \tag{12.1}$$

The most familiar linear equating methods are Tucker equating, Levine equating, and chained linear equating.

In this section, the equations for the Tucker, Levine, and chained linear equating methods will be derived, with details, in a way that shows clearly how they are similar and how they are different, although many results have been published before in the same or slightly different forms (Kolen & Brennan, 2004; von Davier & Kong, 2005; von Davier, 2008; Kane et al., 2009). There are two types of Levine equating, but they are similar in many important ways. In this chapter, they will be considered as a single method until it becomes necessary to distinguish between them.

12.3.1 Chained Linear Equating

The main assumption of chained linear equating is that the symmetric linear links from score X to score A and from score Y to score A are population invariant. The symmetric linear link from X to A in population P is

$$a = f(x) = \alpha_P + \beta_P \mathrm{x}, \tag{12.2}$$

where a is a value of A and x is a value of X, and

$$\beta_P = \sigma_P(A)/\sigma_P(X). \tag{12.3}$$

$$\alpha_P = \mu_P(A) - \beta_P \mu_P(X), \tag{12.4}$$

Under the assumption that Equation 12.2 is population invariant, the terms in Equations 12.3 and 12.4 have the same values in population ***P*** as in population ***S***. Hence we have

$$\beta_P = \sigma_S(A)/\sigma_S(X) = \sigma_P(A)/\sigma_P(X); \tag{12.5}$$

$$\mu_S(A) - \beta_P\mu_S(X) = \mu_P(A) - \beta_P\mu_P(X). \tag{12.6}$$

Then Equations 12.5 and 12.6 can be solved for $\mu_S(X)$ and $\sigma_S(X)$. A similar development leads to formulas for $\mu_S(Y)$ and $\sigma_S(Y)$. Using the identity $\mu_S(A) = w\mu_P(A) + (1 - w)\mu_Q(A)$, under the population invariance assumptions of chained linear equating, the means and standard deviations of X and Y in population ***S*** are given by Equation 12.7:

$$\begin{aligned}
&\text{(a) } \mu_S(X) = \mu_P(X) - (1 - w)[\sigma_P(X)/\sigma_P(A)][\mu_P(A) - \mu_Q(A)],\\
&\text{(b) } \mu_S(Y) = \mu_Q(Y) + w[\sigma_Q(Y)/\sigma_Q(A)][\mu_P(A) - \mu_Q(A)].\\
&\text{(c) } \sigma_S(X) = \sigma_S(A)[\sigma_P(X)/\sigma_P(A)]\\
&\text{(d) } \sigma_S(Y) = \sigma_S(A)[\sigma_Q(Y)/\sigma_Q(A)]
\end{aligned} \tag{12.7}$$

Substituting the terms from Equation 12.7 into Equation 12.1, we get the usual form of the equation for chained linear equating:

$$y = f(x) = \mu_Q(Y) + \frac{\sigma_Q(Y)\sigma_P(A)}{\sigma_Q(A)\sigma_P(X)}[x - \mu_P(X)] + \frac{\sigma_Q(Y)}{\sigma_Q(A)}\big[\mu_P(A) - \mu_Q(A)\big]. \tag{12.8}$$

Notice that the weight w cancels out of Equation 12.8. Chained linear equating does not depend on the relative proportions of populations ***P*** and ***Q*** in population ***S***.

12.3.2 Tucker Equating

The main assumption of Tucker equating is that the regressions of X and Y on A (i.e., the best linear predictors of X and Y from A) are population invariant. The best linear predictor of score X from score A in population ***P*** can be expressed as

$$x = \mathrm{f}'(a) = \alpha'_P + \beta'_P a, \tag{12.9}$$

where x is a value of X and a is a value of A, and

$$\beta'_P = \rho_P(X, A)[\sigma_P(X)/\sigma_P(A)] \tag{12.10}$$

$$\alpha'_P = \mu_P(X) - \beta'_P\mu_P(A) \tag{12.11}$$

These are the values that minimize $\sum \left(X - \alpha'_P - \beta'_P A\right)^2$ in population P.

The assumption that the best linear predictor is population invariant implies that $\beta'_S = \beta'_P$, so that

$$\rho_S(X,A)[\sigma_S(X)/\sigma_S(A)] = \rho_P(X,A)[\sigma_P(X)/\sigma_P(A)], \tag{12.12}$$

which can be solved for $\sigma_S(X)$. The population invariance assumption also implies that $\alpha'_S = \alpha'_P$, so that

$$\mu_S(X) - \beta'_S\mu_S(A) = \mu_P(X) - \beta'_P\mu_P(A). \tag{12.13}$$

Yet, $\beta'_S = \beta'_P$, and $\mu_S(A)$ can be expressed in terms of the known quantities $\mu_P(A)$, $\mu_Q(A)$, and w. Therefore, Equation 12.13 can be solved for $\mu_S(X)$. A similar development leads to formulas for $\mu_S(Y)$ and $\sigma_S(Y)$.

Therefore, under the assumptions that the best linear predictors of X and Y from A are population invariant, the means and standard deviations of X and Y in population S are given by Equation 12.14:

$$\begin{aligned}
&(\text{a})\mu_S(X) = \mu_P(X) - (1-w)\rho_P(X,A)[\sigma_P(X)/\sigma_P(A)][\mu_P(A) - \mu_Q(A)];\\
&(\text{b})\mu_S(Y) = \mu_Q(Y) + w\rho_Q(Y,A)[\sigma_Q(Y)/\sigma_Q(A)][\mu_P(A) - \mu_Q(A)];\\
&(\text{c})\sigma_S(X) = \sigma_S(A)[\rho_P(X,A)/\rho_S(X,A)][\sigma_P(X)/\sigma_P(A)];\\
&(\text{d})\sigma_S(Y) = \sigma_S(A)[\rho_Q(Y,A)/\rho_S(Y,A)][\sigma_Q(Y)/\sigma_Q(A)].
\end{aligned} \tag{12.14}$$

Substituting the estimated means and standard deviations from Equation 12.14 into Equation 12.1, we get this form of the equation for Tucker equating:

$$\begin{aligned}
y = f(x) = \mu_Q(Y) &+ \frac{\rho_Q(Y,A)\rho_S(X,A)\sigma_Q(Y)\sigma_P(A)}{\rho_P(X,A)\rho_S(Y,A)\sigma_Q(A)\sigma_P(X)}[x - \mu_P(X)]\\
&+ \frac{\rho_Q(Y,A)\sigma_Q(Y)}{\sigma_Q(A)}\left[\mu_P(A) - \mu_Q(A)\right]\\
&+ (1-w)\left[1 - \frac{\rho_S(X,A)}{\rho_S(Y,A)}\right]\frac{\rho_Q(Y,A)\sigma_Q(Y)}{\sigma_Q(A)}\left[\mu_P(A) - \mu_Q(A)\right].
\end{aligned} \tag{12.15}$$

The weight w does not cancel out of this equation; Tucker equating depends on the relative proportions of populations P and Q in population S. To compare Tucker equating with chained linear equating, we need to remove this dependence, by finding a realistic condition under which the third term of this expression is zero. In a NEAT design, it is not realistic to assume that populations P and Q have equal mean scores on A. However, it may be realistic to assume that the correlations of X with A and of Y with A are equal in population S: $\rho_S(X, A) = \rho_S(Y, A)$. This assumption leads to the equation for a weight-independent version of Tucker equating:

$$y = f(x) = \mu_Q(Y) + \frac{\rho_Q(Y,A)\sigma_Q(Y)\sigma_P(A)}{\rho_P(X,A)\sigma_Q(A)\sigma_P(X)}[x - \mu_P(X)] + \frac{\rho_Q(Y,A)\sigma_Q(Y)}{\sigma_Q(A)}\left[\mu_P(A) - \mu_Q(A)\right]. \quad (12.16)$$

12.3.3 Levine Equating

The Levine equating methods use the notion of true score from classical test theory, which states that any examinee's score on a test can be decomposed into a *true score*, the part that does not vary over repeated testing, and an *error of measurement*, the part that varies:

$$X = T_X + E_X. \quad (12.17)$$

Errors of measurement are assumed to be purely random — uncorrelated with each other and with true scores — and to have a mean of zero. It follows from this assumption that the correlation of X with T_X in a population is the ratio of their standard deviations in that population:

$$\rho_P(X, T_X) = \sigma_P(T_X)/\sigma_P(X) \quad (12.18)$$

One may see that $\rho_P(X,T_X)$ is simply the square root of the reliability of test X in population P.

For equating test scores through an anchor score, it is possible to make assumptions like those of the Tucker method but, instead of applying them to the observed scores and conditional standard deviations, to apply them to the true scores and standard errors of measurement. That approach leads to the Levine method (Kolen & Brennan, 2004).

The main assumptions of the Levine equating methods are that true scores on X and Y are perfectly correlated with true scores on A and that the functions linking true scores on X and Y to true scores on A are population invariant. By definition, the linear link from T_X to T_A on P is

$$\tau_a = f(\tau_x) = a''_P + b''_P\tau_x, \quad (12.19)$$

where τ_a is a value of T_A and τ_x is a value of T_X, and

$$\beta''_P = \sigma_P(T_A)/\sigma_P(T_X) \quad (12.20)$$

$$\alpha''_P = \mu_P(T_A) - \beta''_P\mu_P(T_X) = \mu_P(A) - \beta''_P\mu_P(X) \quad (12.21)$$

Here in Equation 12.21, the assumption that the mean of E_A (the measurement error of A) is 0 is used to replace the mean of T_A (the true score of A) by the mean of A. Similarly, the mean of T_X is replaced by the mean of X.

Using the population invariance assumption,

$$\beta_P'' = \sigma_S(T_A)/\sigma_S(T_X) = \sigma_P(T_A)/\sigma_P(T_X) \tag{12.22}$$

$$\mu_S(A) - \beta_P''\mu_S(X) = \mu_P(A) - \beta_P''\mu_P(X) \tag{12.23}$$

Equations 12.22 and 12.23 can be solved to get formulas for $\mu_S(X)$ and $\sigma_S(X)$. A similar development leads to formulas for $\mu_S(Y)$ and $\sigma_S(Y)$. Under the assumptions that the linear links between the true scores are population invariant, the means and standard deviations of X and Y over S are given by Equation 12.24:

$$\begin{aligned}
&\text{(a) } \mu_S(X) = \mu_P(X) - (1-w)[\sigma_P(T_X)/\sigma_P(T_A)][\mu_P(A) - \mu_Q(A)],\\
&\text{(b) } \mu_S(Y) = \mu_Q(Y) + w[\sigma_Q(T_Y)/\sigma_Q(T_A)][\mu_P(A) - \mu_Q(A)].\\
&\text{(c) } \sigma_S(X) = \sigma_S(T_A)/\rho_S(X,T_X)[\sigma_P(T_X)/\sigma_P(T_A)]\\
&\text{(d) } \sigma_S(Y) = \sigma_S(T_A)/\rho_S(Y,T_Y)[\sigma_Q(T_Y)/\sigma_Q(T_A)]
\end{aligned} \tag{12.24}$$

Here the result given in Equation 12.18 has been used to replace $\sigma_S(T_X)$ with $\sigma_S(X)$ and similarly for $\sigma_S(T_Y)$.

Substituting the estimated means and standard deviations from Equation 12.24 into Equation12.1, we get this form of the equation for Levine observed-score equating:

$$\begin{aligned}
y = f(x) = \mu_Q(Y) &+ \frac{\rho_S(X,T_X)\sigma_Q(T_Y)\sigma_P(T_A)}{\rho_S(Y,T_Y)\sigma_Q(T_A)\sigma_P(T_X)}[x - \mu_P(X)]\\
&+ \frac{\sigma_Q(T_Y)}{\sigma_Q(T_A)}\left[\mu_P(A) - \mu_Q(A)\right]\\
&+ (1-w)\left[1 - \frac{\rho_S(X,T_X)}{\rho_S(Y,T_Y)}\right]\frac{\sigma_Q(T_Y)}{\sigma_Q(T_A)}\left[\mu_P(A) - \mu_Q(A)\right].
\end{aligned} \tag{12.25}$$

The weight w does not cancel out of this equation; Levine observed-score equating depends on the relative proportions of populations P and Q in population S. To compare Levine observed-score equating with chained linear equating, we need to remove this dependence, by finding a realistic condition under which the third term of this expression is zero. The only likely possibility is that scores X and Y are equally reliable in population S: $\rho_S(X, T_X) = \rho_S(Y, T_Y)$. This assumption leads to a weight-independent version of Levine observed-score equating:

$$y = f(x)$$
$$= \mu_Q(Y) + \frac{\sigma_Q(T_Y)\sigma_P(T_A)}{\sigma_Q(T_A)\sigma_P(T_X)}[x - \mu_P(X)] + \frac{\sigma_Q(T_Y)}{\sigma_Q(T_A)}[\mu_P(A) - \mu_Q(A)] \quad (12.26)$$

Equation 12.26 is identical to the equation for Levine true-score equating.

Now the equations for chained linear equating (Equation 12.8), the weight-independent version of Tucker equating (Equation 12.16), and the weight-independent version of Levine equating (Equation 12.26) can be compared. Those three equations can all be written as Equation 12.27:

$$y = f(x) = \mu_Q(Y) + \frac{\varphi_Q(Y,A)\frac{\sigma_Q(Y)}{\sigma_Q(A)}}{\varphi_P(X,A)\frac{\sigma_P(X)}{\sigma_P(A)}}[x - \mu_P(X)] + \left[\varphi_Q(Y,A)\frac{\sigma_Q(Y)}{\sigma_Q(A)}\right][\mu_P(A) - \mu_Q(A)]. \quad (12.27)$$

where $\varphi_P(X, A)$ and $\varphi_Q(Y, A)$ are factors determined by a given equating.

A comparison of this general Equation 12.27 with the equations for the three linear methods shows that

- if $\varphi_P(X, A) = \varphi_Q(Y, A) = 1$, then Equation 12.27 is Equation 12.8, the equation for chained linear equating;
- if $\varphi_P(X, A) = \rho_P(X, A)$ and $\varphi_Q(Y, A) = \rho_Q(Y, A)$, then Equation 12.27 is Equation 12.16, the equation for the weight-independent version of Tucker equating; and
- if $\varphi_P(X, A) = \rho_P(X, TX) / \rho_P(A, TA)$ and $\varphi_Q(Y, A) = \rho_Q(Y, T_Y) / \rho_Q(A, T_A)$, then Equation 12.27 is Equation 12.26, the equation for the weight-independent version of Levine equating.

The quantities $\rho_P(X, A)$ and $\rho_Q(Y,A)$ are necessarily less than 1. The quantities $\rho_P(X,T_X) / \rho_P(A.T_A)$ and $\rho_Q(Y, T_Y) / \rho_Q(A, T_A)$ are nearly always greater than 1. Therefore, the three methods are ordered, with chained linear in between Tucker and Levine. This relationship explains why, in practical equating situations where the three methods produce different results, the results of chained linear equating nearly always are in between those of Tucker equating and Levine equating.

Of course, there are infinitely many possible values for $\varphi_P(X, A)$ and $\varphi_Q(Y, A)$. Each possible set of values for these parameters leads to a different linear equating method.

12.4 A Generalized Equating Function

Equation 12.27 is a general expression for linear equating in a NEAT design. It leads to a family of linear equating methods, including the Tucker, Levine, and chained linear methods and infinitely many others. But is there a corresponding family of nonlinear equating methods?

The kernel equating procedure provides a way to answer this question. Kernel equating is not a single equating method; it is a procedure that leads to many possible equating methods. Two versions of the kernel equating procedure can be used in a NEAT design: One follows the logic of poststratification equating, and the other follows the logic of chained equating. The kernel equating procedure for poststratification equating of X to Y through A involves four steps:

1. Pre-Smoothing (optional): Fit a log-linear model to each of the bivariate test-anchor score distributions (X, A) in population P and (Y, A) in population Q. The output of this step is a pair of discrete bivariate distributions that are smoother (less irregular) than those observed.
2. Estimation: use the poststratification equating assumption to estimate the score distributions of X and Y (still discrete) in population S.
3. Continuization: replace the discrete distributions estimated for population $\boldsymbol{S}$ by continuous distributions.
4. Equating: link each value of X to the value of Y that has the same percentile rank in the continuized distributions of X and Y estimated for population S.

The kernel procedure for chained equating involves two separate linkings. The four steps of the procedure are as follows:

1. Pre-Smoothing (optional): Fit a log-linear model to each of the bivariate test-anchor score distributions (X, A) in population P and (Y, A) in population Q. The output of this step is a pair of discrete bivariate distributions that are smoother (less irregular) than those observed.
2. Estimation: estimate the (marginal) score distributions of X and A (still discrete) in population P and the score distributions of Y and A in population Q, respectively.
3. Continuization: replace the four discrete marginal distributions by continuous distributions.
4. The equating requires two steps: (a) linking of X to A (link each value of X to the value of A that has the same percentile rank in the continuized marginal distributions of X and A) and (b) linking of A to Y (link each value of A determined in Step 4a to the value of Y that has the same percentile rank in the continuized marginal distributions of Y and A).

The continuization step requires the user of the procedure to specify a bandwidth parameter that determines how far the continuized distributions can depart from the discrete distributions. Small values of the bandwidth parameter make the continuized distribution closely match the discrete distribution, so that the kernel equating very closely resembles the usual type of equipercentile equating. Large values of the bandwidth parameter make the continuized distributions closely resemble normal distributions, so that the kernel equating is, for all practical purposes, linear. Therefore, any linear equating method that can be closely reproduced by a kernel equating procedure with a large bandwidth has an analogous nonlinear method: that same kernel equating procedure with a small bandwidth.

To find the nonlinear equating method that corresponds to a given linear equating method, all that is necessary is to find a kernel equating procedure that is essentially equivalent to that linear equating method. The kernel procedure for chained equating has the natural connection between chained equipercentile equating and chained linear, while the kernel procedure for poststratification equating with large bandwidth produces the Braun-Holland equating method. However, it has been shown in Braun & Holland (1982) that with additional assumptions, the Braun-Holland equating becomes the Tucker equating method. This establishment will serve as the cornerstone to build the generalized equating function based on poststratification equating.

What makes it possible to get the Levine equating from the kernel equating procedures is a transformation that can be applied to the pre-smoothed bivariate test-anchor distributions. This transformation is called the mean-preserving linear transformation (MPLT). The MPLT has the effect of changing the standard deviations of the two variables, while leaving the means unchanged. The transformation has two parameters, and the right choice of values for those parameters will change the kernel equating procedure, so that instead of being essentially equivalent to one linear equating method, it becomes essentially equivalent to another linear equating method. The two parameters, denoted here as λ_X and v_X, function as multipliers for the standard deviations of the two variables. Using X and A represent the test score and anchor score variables before applying the MPLT and X^* and A^* to represent those variables after applying the MPLT,

$$\begin{aligned} &\text{(a) } \mu(X^*) = \mu(X), \\ &\text{(b) } \mu(A^*) = \mu(A), \\ &\text{(c) } \sigma(X^*) = \lambda_X \sigma(X), \\ &\text{(d) } \sigma(A^*) = v_X \sigma(A). \end{aligned} \tag{12.28}$$

If X and A were continuous variables, this change in the standard deviations could be accomplished by simply transforming the variables X and A. However, X and A represent test scores, which are nearly always discrete variables. Transforming X and A would produce a discrete bivariate distribution in which most of the values of each variable would not be possible scores on the test and anchor. Therefore, it is necessary to find a transformation that changes the standard deviations of X and A while keeping the set of possible values unchanged, by redistributing the probabilities. The distribution of X^* and A^* has exactly the same set of possible values as the distribution of X and A. What changes is the probability associated with each pair of values (x, a). (See Brennan & Lee, 2006, and Wang & Brennan, 2007, for details.)

The MPLT can be inserted into the kernel equating procedure by applying it immediately after the pre-smoothing step. The result is a *generalized equating function* with two sets of parameters: a set of four MPLT parameters (λ_X, v_X, λ_Y and v_Y) and a set of bandwidth parameters for the continuization step. Poststratification equating has two bandwidth parameters (which generally have the same value);

chained equating has four bandwidth parameters. Because there are two versions of kernel equating, there are two generalized equating functions, generalized post-stratification equating and generalized chained equating.

Although any set of values for the four MPLT parameters will result in an equating function, some sets of values are better than others—more interesting theoretically and more useful practically. The MPLT makes it possible to change one linear equating into another, by expressing the linear equating as a kernel equating procedure with a large bandwidth and by applying the MPLT to the pre-smoothed test-anchor distributions. In particular, it is possible to find a set of MPLT parameters that will transform the kernel equating procedure that replicates Levine equating (Chen & Holland, 2009). In this case, the MPLT parameters are

$$\begin{aligned} \lambda_X &= 1 + \varepsilon_X, \\ \nu_X &= \lambda_X \rho_P(X,A)\rho_P(A,T_A)/\rho_P(X,T_X), \\ \lambda_Y &= 1 + \varepsilon_Y, \\ \nu_Y &= \lambda_Y \rho_Q(Y,A)\rho_Q(A,T_A)/\rho_Q(Y,T_Y), \end{aligned} \tag{12.29}$$

where ε_X and ε_Y are, respectively, functions of $\rho_P(X, A)\rho_P(A, T_A)\ /\rho_P(X, T_X)$ and $\rho_Q(Y, A)\rho_Q(A, T_A)\ /\rho_Q(Y, T_Y)$, and the values of both are almost zero.

Then, if the large bandwidth parameter in the continuization step is replaced with a small bandwidth parameter, the result is a kernel equating procedure that produces a nonlinear analogue to Levine equating, called *Levine observed-score equipercentile equating*.

Chen and Holland (2009) generalized this procedure by defining a family of MPLT parameters that depend on a single parameter κ as follows:

$$\begin{aligned} \alpha(\kappa) &= [\rho_P(X,A)\rho_P(A,T_A)/\rho_P(X,T_X)]\kappa; \\ \beta(\kappa) &= [\rho_Q(Y,A)\rho_Q(A,T_A)/\rho_Q(Y,T_Y)]\kappa. \end{aligned} \tag{12.30}$$

$$\lambda_X = 1 + \varepsilon_X(\alpha(\kappa)); \nu_X = \lambda_X\alpha(\kappa); \lambda_Y = 1 + \varepsilon_Y(\alpha(\kappa)); \quad \text{and} \quad \nu_Y = \lambda_Y\alpha(\kappa). \tag{12.31}$$

The parameter κ can be any number but preferably should be in the range of [0, 1]. The functions $\varepsilon_X(\,)$ and $\varepsilon_Y(\,)$ have very complicated forms but usually can be ignored when they are evaluated near 1, and $\varepsilon_X(1) = \varepsilon_Y(1) = 0$. This set of MPLT parameters, used in a kernel equating procedure with a small bandwidth, leads to an equipercentile equating associated with κ. Used in the kernel equating procedure for poststratification equating with a large bandwidth, they lead to a linear equating whose weight-independent version is given by Equation 12.27, with $\varphi_Q(Y, A) = \rho_Q(Y, A)^{1-\kappa}(\rho_Q[Y, T_Y]/\rho_Q[A, T_A])^{\kappa}$, and so on. If $\kappa = 0$, then it leads to the Tucker equating; if $\kappa = 1$, then it leads to the Levine equating. Chained linear equating also can be reproduced by an appropriate value for κ, if the correlations of X with A and

of Y with A in population $\boldsymbol{S}$ are nearly equal. In that case, we need to determine κ to make both $\varphi_P(X, A) = 1$ and $\varphi_Q(Y, A) = 1$. To make $\varphi_P(X, A) = 1$,

$$\kappa = \frac{\ln \rho_P(X,A)}{\ln \rho_P(X,A) + \ln \rho_P(A,T_A) - \ln \rho_P(X,T_X)} \tag{12.32a}$$

and to make $\varphi_Q(Y, A) = 1$,

$$\kappa = \frac{\ln \rho_Q(Y,A)}{\ln \rho_Q(Y,A) + \ln \rho_Q(A,T_A) - \ln \rho_Q(Y,T_Y)} \tag{12.32b}$$

If the numbers from both Equations 12.32a and 12.32b are nearly equal, their average can be used as the value of κ, so that $\varphi_Q(Y, A) \approx 1$ and $\varphi_P(X, A) \approx 1$, and the equating function will be very nearly equivalent to chained linear equating. However, because it is derived from poststratification equating, it will be weight dependent.

The special case of generalized poststratification equating in which the MPLT parameters are defined as in Equation 12.31 is called κ-PSE. This generalized equating function, with minor adjustments, can approximate all commonly used methods for equating in a NEAT design. Poststratification equating (i.e., frequency-estimation equipercentile equating) is the special form of κ-PSE with $\kappa = 0$ and a small bandwidth. Braun-Holland and Tucker equating is the special form of κ-PSE with $\kappa = 0$ and a large bandwidth. The Levine methods are the special case of κ-PSE with $\kappa = 1$ and a large bandwidth. Chained equipercentile and chained linear equating are the special cases of κ-PSE with the κ value given in Equation12.32- with a small bandwidth for equipercentile equating and a large bandwidth for linear equating. The hybrid Levine method (von Davier, Fournier-Zajac, & Holland 2006b) is similar to the Levine observed-score equipercentile equating we defined in this section, since both have Levine observed-score equating as their linear form under kernel equating. Chen and Holland (2009) showed that the modified poststratification equating (Wang & Brennan, 2007) for NEAT designs with an external anchor is almost same as the special case of κ-PSE with $\kappa = ½$ and with a small bandwidth. Finally, the chained true-score equipercentile equating developed in Chen and Holland (2008) is the weight-independent version of the Levine observed-score equipercentile equating.

Similarly, we can create a κ-generalized chained equating (κ-CE). The κ-indexed family of λ_X, ν_X, λ_Y, and ν_Y is

$$\begin{aligned} &\lambda_X(\kappa) = \rho_P(X,T_X)^\kappa; \nu_X(\kappa) = \rho_P(A,T_A)^\kappa; \lambda_Y(\kappa) = \rho_Q(Y,T_Y)^\kappa; \quad \text{and} \\ &\nu_Y(\kappa) = \rho_Q(A,T_A)^\kappa. \end{aligned} \tag{12.33}$$

For $\kappa = 0$ with a small bandwidth, the equating is chained equipercentile; with a large bandwidth, it is chained linear. For $\kappa = 1$ with a small bandwidth, the equating

is chained true-score equipercentile equating; with a large bandwidth, it is Levine true-score equating. We can also approximate the poststratification equating and the Tucker equating, but the value of κ will be negative. Notice that κ-CE is weight independent.

12.5 Examples and Discussion

In this section, simulated data based on a data set from an operational testing program is used to demonstrate the generalized equating function, particularly the κ-PSE, illustrating the following specific relationships:

- Tucker equating can be approximated closely by κ-PSE with $\kappa = 0$ and a large kernel equating bandwidth;
- Levine observed-score equating can be approximated closely by κ-PSE with $\kappa = 1$ and a large kernel equating bandwidth;
- Chained equipercentile equating can be approximated closely by κ-PSE with the κ value determined by Equation 12.32and a small kernel equating bandwidth.

The data set contains 2 simulated bivariate distributions, each derived from the same named pre-smoothed distribution described in details in Chapter 10 of A. A. von Davier et al. (2004b). Test X and test Y each contain 78 items; the external anchor A contains 35 items. Table 12.1 shows summary statistics for this simulated data set.

The score distributions in population $\boldsymbol{Q}$ were strongly skewed in a positive direction, on both test form Y and the anchor. In population $\boldsymbol{P}$, the distributions of scores on test form X was skewed, but less strongly, and in the opposite direction.

The first comparison is between the Tucker equating function with $w = 0.5$ and the κPSE with $\kappa = 0$, $w = 0.5$, and kernel equating bandwidth = 5,400. The difference between these two equating functions is less than 0.02 raw-score points (0.0012 *SD*) at all points of the score scale (see Figure 12.1).

The second comparison is between the Levine observed-score equating function with $w = 0.5$ and the κ PSE with $\kappa = 1$, $w = 0.5$, and kernel equating bandwidth = 5,400. The difference between these two equating functions is less than 0.06 raw-score points (0.0036 *SD*) at all points of the score scale (see Figure 12.1).

Table 12.1 Summary Statistics for the Data Set

Statistic	Sample from population P, $n = 10{,}000$		Sample from population Q, $n = 10{,}000$	
	Test X	Anchor A	Test Y	Anchor A
Mean	39.3	17.1	32.5	14.3
SD	17.2	8.4	16.7	8.2
Correlation	0.88		0.88	
Skewness	−0.11	−0.02	0.23	0.26
Kurtosis	2.24	2.15	2.28	2.25

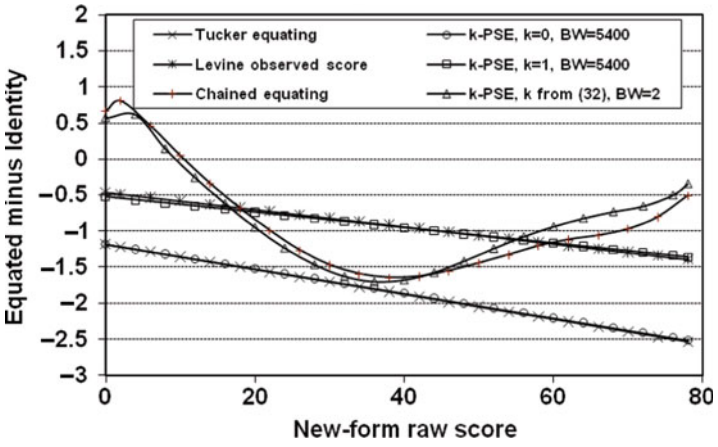

Fig. 12.1 Plots of three equating functions and their counterparts generated by κ-PSE. BW = bandwidth

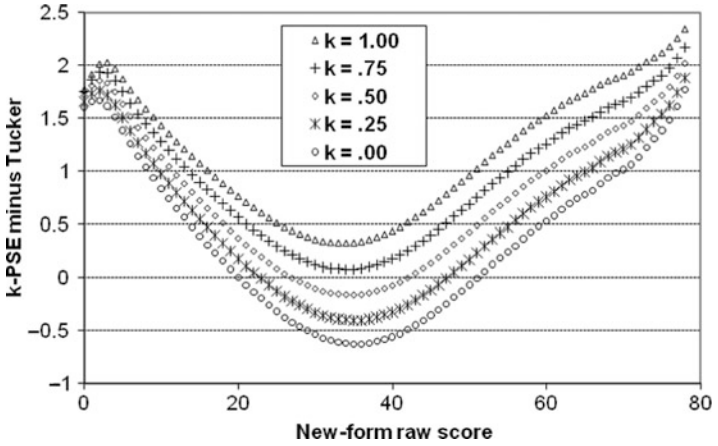

Fig. 12.2 Differences between κ-PSE and Tucker equating, for five values of κ

The third comparison is between the chained equipercentile equating function (produced from kernel equating software with bandwidth = 2) and the κ PSE with κ determined by Equation 12.32 and kernel equating bandwidth = 2. The difference is not as tiny in this comparison as in the previous two comparisons, but still not large; it varies over the score scale from approximately −0.11 to +0.27 raw-score points (−0.007 to +0.016 SD). If both the weight w and the κ were adjusted slightly, the difference would be consistently less than 0.20 raw-score points (0.012 *SD*).

For many years, psychometricians comparing different equating functions computed from the same operational data have observed predictable differences between the Tucker, chained linear, and Levine observed-score equating methods.

When the new-form equating sample scores higher than the reference-form equating sample on the anchor test, the Levine method yields the highest equated scores; the Tucker method yields the lowest. Figure 12.2 shows that similar differences occur with curvilinear κ PSE equating. The plot shows the differences between κ-PSE equating and Tucker equating, for five different values of κ. With $\kappa = 0$ (the lowest curve in the figure), κ-PSE equating corresponds to the usual frequency-estimation method, which is the curvilinear analog to Tucker equating. With $\kappa = 1$ (the highest curve), the κ-PSE equating becomes the curvilinear analog to Levine observed-score equating. As the value of κ increases from 0 to 1, the equated scores become progressively higher.

The kernel equating procedure incorporating the MPLT, with no restrictions on the MPLT parameters λ_X, v_X, λ_Y, and v_X, except that they are all positive, defines a generalized equating function. This generalized equating function provides a framework for creating new equating methods with desired properties. On the other hand, the generalized equating function with κ gives a much better solution for operational work. Each not only pairs the three most familiar linear equatings with its nonlinear counterpart but also expands to a system of equatings indexed with continuous parameters, which the users can choose to get an optimal equating solution based on any criteria they choose.

There are some computation issues. The first is how to compute the true score coefficients defined in Equation 12.18. Currently, we used the formulas in Kolen and Brennan (2004). Interestingly, for NEAT designs using an external anchor, $\alpha(\kappa)$, defined in Equation 12.30 is $\rho_P(A, T_A)^{2\kappa}$; for NEAT designs using an internal anchor, $\alpha(\kappa)$ is $\rho_P(X, A)^{2\kappa}$. These formulas assume that the true scores on the test and the anchor have a perfect linear correlation, which is possible only if the equating relationship is linear. However, even when the relationship is not linear, the correlation of true scores is often close to 1.00 (Chen & Holland, 2008). For most cases, the linear assumption can be used for the computation. More extreme cases are discussed in the Chen and Holland (2008) paper, and the formulas are modified accordingly.

The second issue is how to compute the distribution defined by MPLT (Equation 12.28) on the integers. The distributions of the anchor scores in both samples (from populations $\boldsymbol{P}$ and $\boldsymbol{Q}$) have distributions on the same score points—the possible scores on the anchor. The conditional distributions are computed at these values of the anchor score A. However, the MPLT defined by Equation 12.28 misaligns the anchor scores. Therefore, it is necessary to redistribute the score frequencies at each noninteger value of A to the adjacent integers. The method used by Brennan and Lee (2006) and by Wang and Brennan (2007) produced frequencies of zero at some anchor score points, which distorted the score estimation and made the computation for estimating the standard error of equating impossible. A new method has been created to solve this problem by doing the redistribution in the log-linear pre-smoothing (Chen & Holland, 2010). The implementation of this solution in the pre-smoothing software is currently under development.

12.6 Conclusion

The generalized equating function is built with two basic elements: a base equating—either poststratification equating or chained equating—and the modified kernel equating framework, including the MPLT. If the base equating is poststratification equating, the generalized equating function is called generalized poststratification equating. The generalized poststratification equating is weight dependent; it depends on the relative weights of the two separately sampled examinee populations ($\boldsymbol{P}$ and $\boldsymbol{Q}$) in the combined population for which the equating function is to be estimated. If the base equating is chained equating, the generalized equating function is called generalized chained equating and is weight independent. In some cases, the generalized poststratification equating is not sensitive to differences in the weights, and in those cases, generalized poststratification equating and generalized chained equating are equivalent. Therefore, generalized poststratification equating can be considered the more general approach.

The κ-equating is a special case of the generalized equating function in which the differences between equating methods are expressed as differences in the value of a parameter, called κ. The κ-equating can unite all the commonly used classical methods for equating in a NEAT design, by reducing the selection of each equating to a choice of a value for κ. By expanding the choice to include other values of κ, the κ-equating can be made to generate a whole family of well-defined equating functions. This modification gives equating practitioners a wide choice of available methods and makes it easy to find the equating method that optimizes some specified criterion.

The approach to equating described in this chapter could lead to at least two types of future development and research. One is the development of criteria for the quality of an equating—criteria for choosing among the many possible values of the MPLT and bandwidth parameters. Another is to expand the family of generalized equating functions, by adapting the generalized equating function to include other existing equating methods (e.g., methods based on pre-smoothing the score distributions using item response theory) or by varying parameters of the generalized equating function to create new equating methods.

Author Note: Any opinions expressed in this chapter are those of the authors and not necessarily of Educational Testing Service.