Chapter 19
Hypothesis Testing of Equating Differences in the Kernel Equating Framework

Frank Rijmen, Yanxuan Qu, and Alina A. von Davier

19.1 Introduction

Test equating methods are used to produce scores that are interchangeable across different test forms (Kolen & Brennan, 2004). In practice, often more than one equating method is applied to the data stemming from a particular test administration. If differences in estimated equating functions are observed, the question arises as to whether these differences reflect real differences in the underlying "true" equating functions or merely reflect sampling error. That is, are observed differences in equating functions statistically significant?

By dividing the squared estimated equating difference at a given score point by the square of its asymptotic standard error, a Wald test (Wald, 1943) is obtained to test for the statistical significance of the equating difference. Carrying out a Wald test at a particular score point is tantamount to the decision rule that was proposed by von Davier, Holland, and Thayer (2004b), who used twice the standard error of the equating difference as a critical value for determining whether there is a difference between two equation functions at a given score point (using 1.96 times the standard error would be formally equivalent to carrying out a Wald test at a type I error rate of .05).

Typically, one will be interested in whether one equating function results in different equated scores than another equation function over a range of score points (e.g., a range of potential cut points in a licensure examination). The procedure proposed by von Davier et al. (2004b), being equivalent with carrying out a Wald test at each individual score point with a type I error rate of α, suffers from the multiple testing problem. That is, by carrying out multiple tests at a specific level of significance α, one test for each score point of interest, the actual type I error rate is higher than α. Because the tests are not independent, correcting the significance

F. Rijmen (✉), Y. Qu, and A.A. von Davier
Educational Testing Service, Rosedale Rd., Princeton, New Jersey 08541, USA
e-mail: frijmen@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_19, © Springer Science+Business Media, LLC 2011

level of the individual test by dividing α by the number of tests carried out (the Bonferroni method of adjustment) will be overly conservative.

In this study, we generalize the expressions for the standard error of an equated score and for the standard error of the equating difference at an individual score point to expressions for the variance-covariance matrix of the set of equated scores, and for the variance-covariance matrix of the differences between equating functions over the whole range of score points. The latter matrix can be used to construct a multivariate Wald test for a set of linear functions of differences between equated scores.

The multivariate Wald test offers two main advantages. First, the test can be used as an omnibus test due to its multivariate nature. This way, a set of hypotheses can be tested simultaneously at a type I error rate of α, thus alleviating the multiple testing problem. For example, the Wald test allows for testing the joint hypothesis that there are no differences between two equating functions over a certain range of score points. Second, one can test for a larger variety of hypotheses because each hypothesis is specified as a *linear function* of differences between equating functions. For example, one can test whether, over a certain range of interest, one equating function results in a higher average equated score than another function.

The expressions are derived within the general framework of the kernel method of test equating (von Davier et al., 2004b). In the next section, the kernel method of equating is introduced, together with some notational conventions. Subsequently, we derive the asymptotic variance-covariance matrix for the set of equated scores and for the set of differences between equated scores. The derivation is very similar to the derivation of the asymptotic standard errors presented in von Davier et al. (2004b). In a fourth section, we explain how Wald tests can be constructed based on the expression for the asymptotic variance-covariance matrix of differences in equated scores. The use of this Wald test is illustrated with a dataset from a professional licensure examination.

19.2 The Kernel Method of Equating

The kernel method of equating is a general procedure to equate one test form to another. The kernel method of equating can be described in five steps. The interested reader is referred to von Davier et al. (2004b). Throughout, we adopt the convention that the form to be equated is denoted by Y, and the base form is denoted by X. X and Y also denote the random variables for the score on the respective forms. Without loss of generality, we assume that possible scores on both X and Y range from 1 to J. The following is a brief description of each of the five steps of kernel equating.

1. In a first step, a statistical model for the observed test scores is constructed. Depending on the design, the score distributions of X and Y are modeled separately (equivalent-groups design), their bivariate distribution is modeled (single-group

design, counterbalanced design), or the bivariate distributions of X and an anchor test Z, and of Y and Z are modeled (nonequivalent groups with anchor test design). The score distribution being discrete, a typical choice is to model the score distributions with a log-linear model, but other choices could be made.

The structural part of the statistical model specifies a functional relation between the model parameters and the probabilities of the score distribution. Collecting all these probabilities in a vector **p** and all model parameters in a vector $\boldsymbol{\theta}$,

$$\mathbf{p} = g(\boldsymbol{\theta}). \tag{19.1}$$

That is, the statistical model is a vector-valued function, where each component specifies the probability of a score (or combination of scores when bivariate score distributions are modeled) as a function of the parameters $\boldsymbol{\theta}$ of the statistical model.

2. Second, the probabilities of the scores of X and Y in the target population are expressed as a function of the probabilities obtained in Step 1. The function is called the design function (*DF*) because its form is determined by the test equating design,

$$\mathbf{r} = DF(\mathbf{p}), \tag{19.2}$$

where **r** consists of two subvectors $\mathbf{r}_X$ and $\mathbf{r}_Y$, denoting the score probabilities of X and Y, respectively.

3. Since X and Y are discrete variables, their cumulative distributions F_X and F_Y are piecewise constant functions. In this step, the continuous random variables X_c and Y_c are defined so that their distribution functions F_{sX} and F_{sY} are smooth continuous approximations of F_X and F_Y, respectively. Several smoothing methods are available. The kernel method of equating is named from the use of Gaussian kernel smoothing techniques. The motivation for this smoothing step is that piecewise constant functions are not invertible. However, invertible functions are needed for computing the equipercentile equating function, as given in Equation 19.3. Von Davier et al. (2004b) described how linear and percentile rank equating functions can be mimicked by controlling the smoothness of the functions through a particular choice of the "bandwidth" of a Gaussian kernel. Given a choice for the bandwidth, F_{sX} and F_{sY} functionally depend on **r** only. More specifically, F_{sX} and F_{sY} are weighted sums of J Gaussian cumulative distribution functions, where the weights are the score probabilities collected in $\mathbf{r}_\mathrm{X}$ and $\mathbf{r}_\mathrm{Y}$, respectively (for technical details, see A.A. von Davier et al., 2004b, Chapters 3 & 4).
4. The previous three steps were preparatory steps for the equating step. Here, each possible score on Y is equated to a comparable score on X through the equipercentile equating function,

$$e_X^*(y) = F_{sX}^{-1}(F_{sY}(y)) \tag{19.3}$$

Even though F_{sX} and F_{sY} are functions with a continuous domain, they are only evaluated at the discrete points $y = 1, \ldots, J$, and $F_{sY}(y)$, respectively.

5. Calculating the standard error of equating is the final step. The next section is devoted to this step.

Steps 1–4 are described in a very general way. They will be instantiated differently depending on the equating design, the choice of a statistical model, the desired degree of smoothing, and so forth. Using this quite general framework offers the advantage of demonstrating what is common to many equating methods. At its very general level, the kernel equating method can be described as a vector-valued function with J components, where each component maps score j, $j = 1, \ldots, J$, on Y onto its equated score on X, and where each component is a function of the parameters of the statistical model for the score distribution. Furthermore, this function is a composition of functions itself, reflecting Steps 1–4 of the kernel equating method described above. Let **eq** denote the vector-valued function that describes Steps 1–4,

$$\mathbf{eq} = \begin{pmatrix} eq_1(\boldsymbol{\theta}) \\ \vdots \\ eq_j(\boldsymbol{\theta}) \\ \vdots \\ eq_J(\boldsymbol{\theta}) \end{pmatrix}. \tag{19.4}$$

Each component j of **eq** can be written as a composition of the functions described in Steps 1–4:

$$eq_j = e_X^j \circ DF \circ g, \tag{19.5}$$

Hence, starting with model parameters $\boldsymbol{\theta}$, the score probabilities $\mathbf{r}$ of X and Y are obtained by applying the design function DF to $\mathbf{p} = g(\boldsymbol{\theta})$. The score probabilities provide the weights for the smoothed cumulative distribution functions F_{sX} and F_{sY}, so that e_X^j is a function that maps $\mathbf{r}$ on the equated score on X of the j^{th} score on Y. Note that there are J e_X^j functions, one for each score Y. Furthermore, e_X^* in Equation 19.3 is a function that maps y onto its equated score on X, whereas e_X^j maps $\mathbf{r}$ on the equated score on X.

19.3 The Standard Error of Equating

Population equating functions are estimated from a sample and therefore subject to sampling variability. The standard error is a measure of the variability of the estimated quantities. Von Davier et al. (2004b) described how the standard errors of equated scores can be obtained using the delta method. Without going into the mathematical details (which can be found in, e.g., Lehmann, 1999), the delta method is based on the property that, if a vector of parameter estimates $\hat{\boldsymbol{\beta}}$ is (asymptotically) normally distributed with variance matrix $\hat{\mathbf{I}}$, a vector-valued continuously differentiable function f of $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed as well, and its variance is obtained by pre- and postmultiplying $\hat{\mathbf{I}}$ with the Jacobian matrix $\mathbf{J}_f$ of the function evaluated at the parameter estimates,

$$\mathrm{COV}(f(\hat{\boldsymbol{\beta}})) = \mathbf{J}_f(\hat{\boldsymbol{\beta}})\Sigma(\mathbf{J}_f(\hat{\boldsymbol{\beta}}))^t \tag{19.6}$$

The delta method is based on a first-order Taylor approximation of f in $\hat{\boldsymbol{\beta}}$, and therefore, for a finite sample size, the asymptotic approximation will be less accurate the more nonlinear f is in the neighborhood of $\hat{\boldsymbol{\beta}}$. The rank of the covariance matrix $\mathrm{COV}(f(\hat{\boldsymbol{\beta}}))$ is at most the minimum of the ranks of $\hat{\mathbf{I}}$ and $\mathbf{J}_f$. Hence, a necessary condition to ensure that the distribution of $f(\hat{\boldsymbol{\beta}})$ is a proper distribution is that the dimensionality of $f(\hat{\boldsymbol{\beta}})$ is not larger than the dimensionality of $\hat{\boldsymbol{\beta}}$.

Applied to the kernel method of equating, the parameters of the statistical model for the score distributions, $\boldsymbol{\theta}$, play the role of $\boldsymbol{\beta}$ in Equation 19.6, and the equation function eq the role of function f. Hence,

$$\mathrm{COV}(\mathbf{eq}(\hat{\boldsymbol{\theta}})) = \mathbf{J}_{eq}(\hat{\boldsymbol{\theta}})\Sigma(\mathbf{J}_{eq}(\hat{\boldsymbol{\theta}}))^t \tag{19.7}$$

Since eq is a composition of functions, its Jacobian may be computed as the product of their Jacobians (the chain rule of differentiation):

$$\begin{aligned}\mathbf{J}_{eq} &= \frac{\partial\, eq(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\ &= \frac{\partial\, e_X(\mathbf{r})}{\partial\mathbf{r}}\,\frac{\partial DF(\mathbf{p})}{\partial\mathbf{p}}\,\frac{\partial g(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\ &= \mathbf{J}_e\mathbf{J}_{DF}\mathbf{J}_g\end{aligned} \tag{19.8}$$

Holland and Thayer (1987) gave expressions for $\mathbf{J}_g$ when the score distributions are modeled with log-linear models, and von Davier et al. (2004b) presented $\mathbf{J}_{DF}$ for the different equating designs. Von Davier et al. also gave the row vector of derivatives $(\frac{\partial e_X^j}{\partial\mathbf{r}})^t$, which forms the j$^{\text{th}}$ row of $\mathbf{J}_e$ in Equation 19.8.

The rank of the variance matrix $\mathrm{COV}(\mathbf{eq}(\hat{\boldsymbol{\theta}}))$ is at most the minimum of the ranks of $\mathbf{J}_e$, $\mathbf{J}_{DF}$ and $\mathbf{J}_g$. Hence, unless a completely saturated log-linear model is used during presmoothing, $\mathrm{COV}(\mathbf{eq}(\hat{\boldsymbol{\theta}}))$ will not be of full rank, and the multivariate distribution of $\mathbf{eq}(\hat{\boldsymbol{\theta}})$ is degenerate. However, Equation 19.7 is still useful, since the asymptotic distributions of single equated scores and pairs, triples, and so forth of equated scores are simply obtained by selecting the corresponding entries in $\mathrm{COV}(\mathbf{eq}(\hat{\boldsymbol{\theta}}))$.

Von Davier et al. (2004b) also presented expressions for the standard error of the difference between two equating functions evaluated in the same score of Y. Using Equations 19.5 and 19.7, their result is easily generalized to the variance-covariance matrix of the vector of differences between two equating functions. In particular, let the same log-linear model be used for both equating functions. Having the same design function by definition, the equating difference function mapping each score of Y into the difference of equated scores is a vector-valued function with as the j^{th} component

$$\Delta^j_{\mathbf{eq}} = \left(e^j_{1X} - e^j_{2X}\right) \circ DF \circ g, \tag{19.9}$$

The asymptotic variance matrix is obtained as

$$\mathrm{COV}\left(\Delta_{\mathbf{eq}}\left(\hat{\boldsymbol{\theta}}\right)\right) = \mathbf{J}_{\Delta_{eq}}\left(\hat{\boldsymbol{\theta}}\right)\Sigma\left(\mathbf{J}_{\Delta_{eq}}\left(\hat{\boldsymbol{\theta}}\right)\right)^t, \tag{19.10}$$

where

$$\begin{aligned} \mathbf{J}_{\Delta_{eq}} &= \frac{\partial\, \Delta_{\mathbf{eq}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial\,(e_{1X} - e_{2X})(\mathbf{r})}{\partial \mathbf{r}}\, \frac{\partial DF(\mathbf{R})}{\partial \mathbf{R}}\, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= (\mathbf{J}_{e_1} - \mathbf{J}_{e_2})\mathbf{J}_{DF}\mathbf{J}_g \end{aligned} \tag{19.11}$$

19.4 Wald Tests to Assess the Difference Between Equating Functions

In this section, we present a generalization of the Wald test presented in von Davier et al. (2004b), which tests for the difference between two equating functions at individual score points. Von Davier et al. divided the equating difference at a given score point by its asymptotic standard error. This statistic is asymptotically standard normally distributed under the null hypothesis that there is no difference between the two equating functions. This is a specific instantiation of the Wald test.

In its general form, the Wald statistic to test a set of linear hypotheses $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, where each row of $\mathbf{L}$ represents a linear hypothesis on $\boldsymbol{\beta}$, has the following form:

$$w = \left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{L}\,\mathrm{COV}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{L}'\right)^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}. \tag{19.12}$$

If $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed, w is asymptotically chi-squared distributed with as degrees of freedom the number of rows of $\mathbf{L}$.

In the context of testing for a difference between two equating methods, $\Delta_{eq}(\hat{\boldsymbol{\theta}})$ fulfills the role of $\hat{\boldsymbol{\beta}}$, and its covariance matrix is given in Equations 19.10 and 19.11. Hence,

$$w_{\Delta_{eq}} = (\mathbf{L}\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))'(\mathbf{L}\,\mathrm{COV}(\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))\mathbf{L}')^{-1}\mathbf{L}\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}})\ldots \tag{19.13}$$

Even though $\mathrm{COV}(\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))$ is in general not of full rank, as explained above, a valid test statistic can be obtained as long as $\mathbf{L}\,\mathrm{COV}(\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))\mathbf{L}'$ is of full rank. In general, if the number of linear hypotheses tested is low, this will be the case, because each linear hypothesis represents a row in **L**, and the number of rows in **L** equals the size of $\mathbf{L}\,\mathrm{COV}(\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))\mathbf{L}'$.

Three immediate choices for **L** come to mind. First, if **L** is a vector of zeros except for element *j*, which equals 1, the test developed by von Davier et al. (2004b) for testing for the equating difference at a single score point is obtained. Second, one can test the joint hypothesis that the equating difference is different from zero at a subset of score points. The number of hypotheses that can be tested simultaneously equals the rank of $\mathrm{COV}(\boldsymbol{\Delta}_{\mathbf{eq}}(\hat{\boldsymbol{\theta}}))$ and hence is bounded from above by the number of parameters of the log-linear model. Third, if interest lies only in an average difference on a certain range of scores, this is accomplished by letting **L** be a vector with its j^{th} element equal to one if the equating difference at score *j* is of interest, and zero otherwise.

19.5 Application

The use of the multivariate Wald test is illustrated with data stemming from two forms of a professional licensure test. The data were collected under an equivalent-groups design. The descriptive statistics for both forms are provided in Table 19.1. The test scores ranged from 0 to 40. Cut points for passing the licensure examination ranged from 24 to 32 on the base form. The results in this section are based on a random sample of 1,000 examinees for each form.

Using Akaike's (1974) information criterion as a selection criterion, log-linear models with 6 and 6 moments were selected, for Test Forms X and Y (*X* and *Y*), respectively. Two equating functions were computed. In the first equating function, equipercentile equating, the bandwidths of the Gaussian kernels (used during the continuization step of the kernel method of equating) were automatically chosen by minimizing the sum of the first and second penalty function presented in von Davier et al. (2004b, Equation 4.30, $K = 1$). The optimal bandwidth values

Table 19.1 Descriptive Statistics for the Raw Scores of Forms X and Y

Form	*N*	Mean	*SD*	Min.	Max.	Skew	Kurtosis	Reliability
X	5,407	29.40	7.17	6.00	40.00	−0.68	−0.23	0.88
Y	5,389	30.69	7.24	4.00	40.00	−0.91	0.09	0.90

were 0.54 and 0.53, for form X and form Y, respectively. For the second equating function (linear equating), the bandwidths were set to a large value, 100 times the standard deviation of the scores for each form, in order to mimic the linear equating function. The difference between the equipercentile (with optimal bandwidth) and linear equating function is plotted in Figure 19.1, together with the 95% confidence bands (and 99.4% confidence bands, see below). Since the possible cut points range from 24 to 32, special interest lies in whether the two equating functions are significantly different from each other in that range. For scorepoints 26–32, zero is outside the 95% confidence interval, suggesting that the two equating functions result in a significantly different equated score for those score points.

Because one is testing nine hypotheses, the overall type I error is much higher than 0.05. Applying the Bonferroni correction for multiple testing (Abdi, 2007), the two equating functions are declared significantly different at a score point when zero falls outside the 99.4% (100 – 5/9) confidence interval. In this case, the difference is no longer significant at score points 26 and 27.

As can be seen in Figure 19.1, the difference between the equipercentile and the linear equating function is a smooth function. Hence, testing for a difference between the two equating functions at scorepoint j is likely not independent of the test at scorepoint $j + 1$. Consequently, the Bonferroni correction for multiple testing is too conservative.

Fisher's (1960) protected least significance difference test offers a procedure to control the overall type I error rate while being less conservative than the

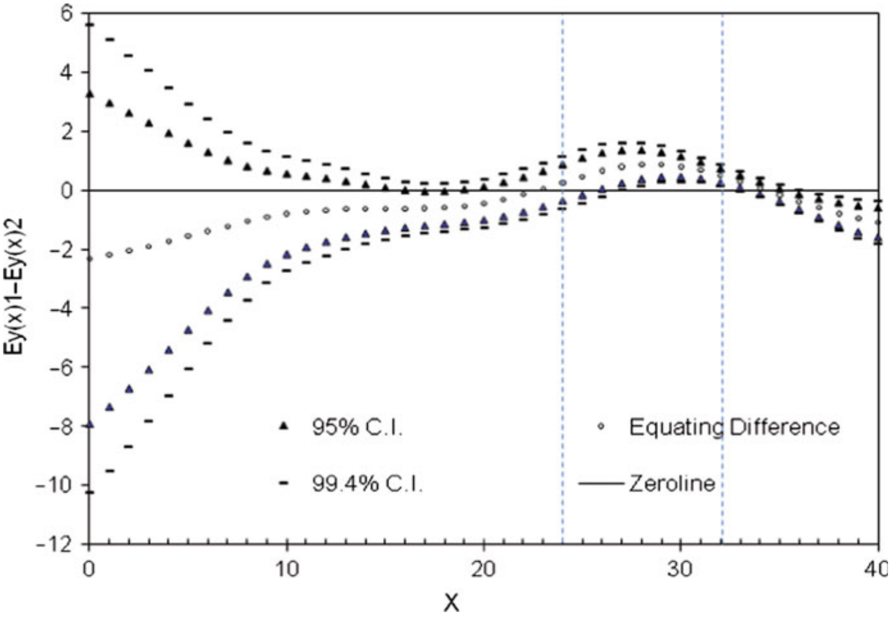

Fig. 19.1 Equating differences and their confidence intervals

Bonferroni correction. In this context, first an omnibus test is carried out at a specific α level that tests the null hypothesis that there is no difference between the two equating functions at any of the nine score points within the range of interest. Only if the omnibus test rejects the null hypothesis, an individual test at the same α level as the omnibus test is carried out for each score point.

The omnibus test was performed using the multivariate Wald test. **L** consisted of nine rows, one for each score point in the range from 24 to 32, where in each row j all entries were set to zero, except entry $23 + j$. The Wald statistic was 7,146.81, $df = 9, p < .001$. Hence, the two equation functions were not the same in the range 24 to 32. Carrying out the second step of Fisher's protected least significant difference test at $\alpha = 0.05$ revealed significant differences for scores 26 to 32. Note that Fisher's protected least squares difference test does not allow for claiming significant differences outside the range from 24 to 32 when zero falls outside the 95% confidence interval. The reason is that those score points were not included in the omnibus test and thus were not "protected" against an inflated type I error rate due to multiple testing.

As a second illustration of the use of the Wald test, we tested whether the equipercentile equating function resulted in higher equated scores on average over the score range from 24 to 32. This hypothesis was tested by defining **L** to be a vector with its j^{th} element equal to one if $24 \leq j \leq 32$, and zero otherwise. The Wald statistic amounted to 7.61, $df = 1$, $p < .001$. On average, the curvilinear equating function resulted in higher equated scores. This means that, on average over all potential cut points, more examinees would pass the test if form Y was equated to form X with a curvilinear equation function than if when a linear equating function were used.

19.6 Discussion

In this paper, we presented the general expressions for the variance-covariance matrix of the differences in equated scores stemming from different equating functions. The derivations were presented within the overall framework of the kernel method of equating (von Davier et al., 2004b). This matrix can be used to construct a multivariate Wald test for a set of linear functions of differences between equated scores.

The multivariate Wald test is general and versatile in its use, as was illustrated with data stemming from a professional licensure exam. A first use is to specify a multivariate omnibus test as a first step in Fisher's protected least significance difference test. The use of the omnibus test protects for an inflated type I error rate due to multiple testing, in that no individual tests are carried out if the omnibus test does not reject the null hypothesis.

In the example, an omnibus test was specified to test for the difference between two equating functions over the range of potential cut scores of the test. The result indicated that the curvilinear equating function did not result in the same equated scores as the linear equating function overall. The tests for differences at the individual score points that were carried out as the second stage of Fisher's least significant difference test indicated that the two equating functions resulted in different equated scores at cut scores 26 to 32.

Fisher's protected least significant difference test is a very old procedure, and many other procedures exist (Carmer & Swanson, 1973; Kuehl, 2000). Insofar these procedures incorporate the use of an omnibus test, the multivariate Wald test will be a part of these procedures as well.

Second, the Wald test can be used to test any hypothesis on a linear combination of differences between equating functions. In the application, this use was illustrated by constructing a test for the average difference between the curvilinear and the linear equating function over the range of potential cut scores. The result indicated that, on average, the curvilinear equating function resulted in higher equated scores than the linear equating function.

The Wald test has many other useful applications. For example, when a test is used to classify examinees in more than two proficiency levels, more than one cut point has to be specified. An omnibus Wald test could be used to test whether two equating functions are different at any of the cut points. In addition, one could test whether one equating function consistently leads to more examinees in higher proficiency levels.

The data of the application were collected using an equivalent-groups design, and subsequently we compared the linear with the equipercentile equating function. However, the expressions for the variance-covariance matrix of the set of equated scores and for the variance-covariance matrix of the differences between equating functions were presented within the general framework of kernel equating. Therefore, multivariate Wald tests can be constructed straightforwardly to assess differences between equating functions other than the linear and equipercentile function, and for data collection designs other than the equivalent-groups design. For example, it may be of interest to use multivariate Wald tests to assess the differences between chain equating and poststratification methods in a nonequivalent groups with anchor test design.

The fact that the difference between two equating functions is statistically significant is only one part of the story. For large samples, even a small difference may reach statistical significance. In order to judge the practical significance between two equating functions, Dorans and Feigenbaum (1994) introduced the difference that matters, the difference that would result in a different score after rounding. A difference between two equating functions is judged to be practically significant whenever it exceeds the difference that matters.

Author Note: The authors would like to thank Tim Moses for sharing SAS macros to carry out kernel equating and to compute the standard error of the equating difference at a given score point. The results reported in this paper were obtained through an adaptation of Tim's SAS macros. Any opinions expressed in this chapter are those of the authors and not necessarily of Educational Testing Service.