Research Report

Using the Kernel Method of Test Equating for Estimating the Standard Errors of Population Invariance Measures

Tim Moses

Research & Development

July 2006
RR-06-20

Using the Kernel Method of Test Equating for Estimating the Standard Errors of Population Invariance Measures

Tim Moses

ETS, Princeton, NJ

July 2006

Abstract

Population invariance is an important requirement of test equating. An equating function is said to be population invariant when the choice of (sub)population used to compute the equating function does not matter. In recent studies, the extent to which equating functions are population invariant is typically addressed in terms of practical significance and not in terms of the equating functions' sampling variability.

This paper shows how to extend the framework of kernel equating to evaluate population invariance in terms of statistical significance. Derivations based on the kernel method's standard error formulas are given for computing the standard errors of the root mean square difference (RMSD) and of the simple difference between two subpopulations' equated scores. An investigation of population invariance for the equivalent groups design is discussed. The accuracy of the derived standard errors is evaluated with respect to empirical standard errors. This evaluation shows that the accuracy of the standard error estimates for the equated score differences is better than for the RMSD and that accuracy for both standard error estimates is best when sample sizes are large.

Key words: Population invariance, equating, standard errors

Acknowledgments

The author thanks Alina von Davier and Paul Holland for discussions and help in several important derivations of the results in this paper and the College Board for permission to use their test data.

Table of Contents

List of Tables

List of Figures

Introduction

Population invariance is an important requirement of test equating, the process of adjusting the scores of test forms so that they are comparable (Angoff, 1971; Dorans & Holland, 2000; Kolen & Brennan, 1995; Lord, 1980; Peterson, Kolen, & Hoover, 1989). An equating function is said to be population invariant when the choice of (sub)population used to compute the equating function does not matter. The extent to which equating functions are population invariant is commonly addressed in terms of *differences that matter* (Dorans, Holland, Thayer, & Tateneni, 2004; Yang, 2004; Yin, Brennan, & Kolen, 2004), that is, the differences in equated scores that are greater than would be corrected by the score rounding that occurs before score reporting (Dorans & Feigenbaum, 1994). The differences-that-matter criterion indicates the practical, rather than the statistical, significance of equating function differences.

The focus of this paper is on incorporating equating functions' sampling variability into population invariance measures in order to evaluate population invariance with respect to statistical significance. The approach taken in this paper is based on the delta method, which is different from other approaches, including those that have conducted significance tests of scale score distribution differences (Segall, 1997), drawn random samples to estimate standard errors of equating functions (Angoff & Cowell, 1986), and utilized bootstrap resampling for item parameter estimates (Williams, Rosa, McLeod, Thissen, & Sanford, 1998). In this paper, derivations are given for computing theoretical standard errors of the root mean square difference (RMSD; Dorans & Holland, 2000) and the standard error of the simple difference between two independent subpopulations' equated scores. These derivations are intended to be applied using the kernel method of observed score equating (von Davier, Holland, & Thayer, 2004; Holland & Thayer, 1989). The kernel method is an equating method with a framework that is general enough to consider population invariance at each explicit step of the equating process, across linear and curvilinear equating functions, and across all of the major equating designs.

The outline of this paper is as follows. First, the steps of kernel equating are reviewed and extended for the computation of subpopulations' and populations' equating functions, population invariance measures, and their standard errors. Next, an example of an investigation of population invariance for the equivalent groups design is given. This example is used to demonstrate and evaluate the proposed standard errors. Finally, the implications of this investigation and applications of the derivations used in this paper to evaluating population invariance for other equating designs are discussed.

Equivalent Groups Kernel Equating and Extensions for Population Invariance Measures

Kernel equating is a unified approach to test equating based on a flexible family of curvilinear and linear equating functions (von Davier et al., 2004; Holland & Thayer, 1989). While the kernel method can be used to compute equating functions based on any of the major data collection designs, the focus of this paper is on linking functions computed from the equivalent groups data collection design (reviewed next). In this section, the five steps of the kernel method are summarized for the equivalent groups design. Extensions of these steps are given for the consideration of population invariance in linking functions that provide (a) a significance test of the difference between two subpopulations' linking functions and (b) the standard error for the RMSD measure.

The Equivalent Groups Design

The equivalent groups design is a data collection design where two independent random samples are drawn from a common population of examinees, *P*, and one sample is administered test *X* and the other sample is administered test *Y*. Table 1 shows the population, samples, and tests administered in the equivalent groups design.

The assumptions of the *X*-to-*Y* equating are the following:

1. There is a single population *P* of examinees who could take either test.
2. The two samples are independently and randomly drawn from the common population of examinees, *P*.

Table 1

The Equivalent Groups Data Collection Design

Population	Sample	Test *X*	Test *Y*
P	1	√	
P	2		√

When the equivalent groups design is extended to a consideration of subpopulations of the samples who take tests *X* and *Y*, the assumptions of the design are also extended. Table 2 shows the equivalent groups design when there are *G* subpopulations making up population *P*. The assumptions made in the equivalent groups design for the *X*-to-*Y* equatings for $P_1,\ldots,P_\mathrm{G}$ adds a third assumption:

3. The *G* subpopulations are mutually exclusive and independent.

Table 2

The Equivalent Groups Data Collection Design for G Subpopulations

Population	Sample and subpopulation	Test *X*	Test *Y*
P	P_1	√	
	P_2	√	
	.	√	
	.	√	
	P_G	√	
P	P_1		√
	P_2		√
	.		√
	.		√
	P_G		√

The assessment of the population invariance equating assumption using population invariance measures is an empirical evaluation based on comparing the *X*-to-*Y* equated scores for $P_1,\ldots,P_G$ and *P*. The five steps of the kernel method and their extensions to consider population invariance are now summarized for the equivalent groups design.

Step 1: Presmoothing, Kernel Equating

Estimates of the univariate score distributions and the *C*-matrices, the factorization matrices of the covariance matrix of the estimated distributions, are obtained by fitting loglinear smoothing models (Holland & Thayer, 1989, 2000) to the raw data obtained by the data collection design. For the equivalent groups design, two loglinear smoothing models are used to preserve a number of moments, T_X and T_Y, from the observed distributions of each separate test,

X and Y, respectively. These loglinear smoothing models produce estimated univariate distributions, $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$, and their corresponding root covariance C-matrices, $\mathbf{C}_{\mathbf{R}}$ and $\mathbf{C}_{\mathbf{S}}$, for X and Y in the total population P. For an equivalent groups design and an X with J possible score values and a Y with K possible score values, the dimensions of $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$ are J-by-1 and K-by-1 and the dimensions of $\mathbf{C}_{\mathbf{R}}$ and $\mathbf{C}_{\mathbf{S}}$ are J-by-T_X and K-by-T_Y.

Population Invariance Measures Extension of Step 1

When kernel equating is extended to a population invariance study, two loglinear smoothing models for each subpopulation's X and Y score distributions are needed to obtain $\hat{\boldsymbol{R}}_{\boldsymbol{Pg}}$ and $\hat{\boldsymbol{S}}_{\boldsymbol{Pg}}$, and $\mathbf{C}_{\mathbf{RPg}}$ and $\mathbf{C}_{\mathbf{SPg}}$ for all G. Since the total population's X and Y distributions are functions of the subpopulations' distributions under the third assumption, smoothing models for the total population are not necessary. The justification of the separate smoothing models and outputs follows from the assumption of independent subpopulations, who consequently do not share the parameters of their respective smoothing models with those of any other subpopulation.

Step 2: Estimation of the Score Probabilities, Kernel Equating

In Step 2, a column vector of estimated score probabilities ($\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{s}}$) is obtained from the estimated score distributions ($\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$) through a design function. For the equivalent groups design, the estimated score probabilities are equal to the estimated score distributions ($\hat{\boldsymbol{r}} = \hat{\boldsymbol{R}}$, $\hat{\boldsymbol{s}} = \hat{\boldsymbol{S}}$, $\mathbf{C}_{\mathbf{r}} = \mathbf{C}_{\mathbf{R}}$, and $\mathbf{C}_{\mathbf{s}} = \mathbf{C}_{\mathbf{S}}$). Therefore, the design function is the identity function and hence, the matrix of derivatives of the design function with respect to the score probabilities, $\mathbf{J}_{\mathbf{DF}}$, is an identity matrix for X's J-total score probabilities ($\boldsymbol{I_J}$) and another identity matrix for Y's K-total score probabilities ($\boldsymbol{I_K}$). For other equating designs, the design function and its derivative matrix ($\mathbf{J}_{\mathbf{DF}}$) involve additional computations for the other equating designs because for other equating designs, $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{s}}$ are not equal to $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$.

Population Invariance Measures Extension of Step 2

For the extension to the population invariance measures, the estimated score probabilities from the smoothed distributions must be produced for each of the G subpopulations and, for the RMSD measure, also for the overall population. For each subpopulation, the score probabilities are estimated from the smoothed distributions in exactly the same way as they are for kernel equating, which is simply through the design function. Once these subpopulation score probabilities are produced, the overall population score probabilities can be estimated as explicit functions of the subpopulation samples and score probabilities (Assumption 3):

$$\hat{r}_{jP} = \frac{\sum_{g}^{G} n_{XPg} \hat{r}_{jPg}}{\sum_{g}^{G} n_{XPg}}, \tag{1}$$

$$\hat{s}_{kP} = \frac{\sum_{g}^{G} n_{YPg} \hat{s}_{kPg}}{\sum_{g}^{G} n_{YPg}}, \tag{2}$$

$$\hat{r}_{jPg} = \text{Prob}\{X = x_j \mid P_g\},$$

$$\hat{s}_{kPg} = \text{Prob}\{Y = y_k \mid P_g\},$$

n_{XPg} is the total number of examinees taking test *X* in subpopulation P_g, and n_{YPg} is the total number of examinees taking test *Y* in subpopulation P_g.

Step 3: Continuation, Kernel Equating

In Step 3, continuous approximations, $\hat{F}_{hX}(x)$ and $\hat{G}_{hY}(y)$, to the estimated discrete cumulative density functions (cdfs), $\hat{F}(x)$ and $\hat{G}(y)$ are determined using Gaussian kernel smoothing (Ramsay, 1991). This step involves a choice of the bandwidth parameters, h_X and h_Y. von Davier et al. (2004) suggested two criteria for selecting the bandwidth parameters: (a) the bandwidth parameters should produce probability density functions that closely match the smoothed discrete probabilities, and (b) the bandwidth parameters should produce probability density functions that do not have many modes.

Population Invariance Measures Extension of Step 3

The continuous cdf approximations $\widehat{F}_{hX}(x)$ and $\widehat{G}_{hY}(y)$ are determined for each subpopulation's cumulative distribution (cumulated Equations 1 and 2) by selecting subpopulation-specific bandwidth parameters h_{XPg} and h_{YPg} for $\widehat{F}_{hXPg}(x)$ and $\widehat{G}_{hYPg}(y)$, and also for the population-specific bandwidth parameters h_{XP} and h_{YP} for $\widehat{F}_{hXP}(x)$ and $\widehat{G}_{hYP}(y)$. Each cdf approximation is based on the corresponding subpopulation or population score probabilities estimated in Step 2.

Step 4: Equating, Kernel Equating

The estimated equating function is formed from the continuous cdfs, $\widehat{F}_{hX}(x)$ and $\widehat{G}_{hY}(y)$, using the following formula: $\widehat{e}_Y(x) = \widehat{G}_{hY}{}^{-1}(\widehat{F}_{hX}(x))$.

Population Invariance Measures Extension of Step 4

Equating functions are computed for each subpopulation as $\widehat{e}_{YPg}(x) = \widehat{G}_{hYPg}{}^{-1}(\widehat{F}_{hXPg}(x))$. The equating function is also computed for the total population as $\widehat{e}_{YP}(x) = \widehat{G}_{hYP}{}^{-1}(\widehat{F}_{hXP}(x))$.

The RMSD can now be computed to measure the extent to which subpopulations' *X-to-Y* equated scores differ from the total population's equated scores at specific values of *X* (x_j, j = *0* to *J*):

$$RMSD(x_j) = \frac{\sqrt{\sum_g w_g [e_{YPg}(x_j) - e_{YP}(x_j)]^2}}{\sigma_{YP}}. \qquad (3)$$

For the RMSD, *g* defines one of the *G* total subpopulations (P_g) of the total population *P* ($\sum_g^G P_g = P$), w_g is the relative proportion of subpopulation P_g in the total population ($w_g = \frac{n_{XPg} + n_{YPg}}{\sum_g^G n_{XPg} + \sum_g^G n_{YPg}}$), $e_{YPg}(x_j)$ is the linking function for *X* to *Y* for a particular score on *X* (x_j) in subpopulation P_g, $e_{YP}(x_j)$ is the linking function for *X*-to-*Y* for score (x_j) in the total population

P, and σ_{YP} is the standard deviation of the *Y* scores in the total population *P*

$$(\sigma_{YP} = \sqrt{\sum_k (y_k - \mu_{YP})^2 \hat{s}_{kP}}).$$

Differences in two subpopulations' equated scores may also be of interest, $\hat{e}_{YP1}(x) - \hat{e}_{YP2}(x)$, particularly if differences in subpopulations' equated scores are regarded as more serious than the differences of subpopulations' equated scores to the total population's equated scores measured by the RMSD.

Step 5: Calculating the Standard Error of Equating, Kernel Equating

The delta method is used to compute a large-sample approximation of the standard error of equating (Bishop, Feinberg, & Holland, 1975; Kendall & Stuart, 1977). The delta method can be summarized by saying that if a vector of parameter estimates ($\hat{\boldsymbol{\theta}}_n$) is distributed approximately as $N(0,\boldsymbol{\Sigma}(\theta))$ when the variance-covariance matrix $\boldsymbol{\Sigma}(\theta)$ is small, then a function of these parameter estimates, $R(\hat{\boldsymbol{\theta}}_n)$, has an approximate $N(0,\partial R/\partial\boldsymbol{\theta}\boldsymbol{\Sigma}(\theta)\partial R/\partial\boldsymbol{\theta}^t)$ distribution (von Davier et al., 2004, p. 198). For kernel equating, the loglinear smoothing output corresponds to ($\hat{\boldsymbol{\theta}}_n$), the *C*-matrix is the factorization of $\boldsymbol{\Sigma}(\theta)$, and the design and equating functions are $R(\hat{\boldsymbol{\theta}}_n)$. Therefore, the standard error of equating reflects the smoothed distributions, the conversion of the smoothed distributions into score probabilities, and the bandwidth-dependent equating functions. For the equivalent groups design:

$$SEE_Y(x) = \left\| \left(\frac{\partial e_Y}{\partial \boldsymbol{r}}, \frac{\partial e_Y}{\partial \boldsymbol{s}} \right) \begin{pmatrix} \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{R}} \mathbf{C_R} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{S}} \mathbf{C_S} \end{pmatrix} \right\| = \left\| \left(\frac{\partial e_Y}{\partial \boldsymbol{r}}, \frac{\partial e_Y}{\partial \boldsymbol{s}} \right) \begin{pmatrix} \mathbf{C_r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_s} \end{pmatrix} \right\|, \tag{4}$$

where $\left(\frac{\partial e_Y}{\partial \boldsymbol{r}}, \frac{\partial e_Y}{\partial \boldsymbol{s}} \right)$ are the partial derivatives of the equating function with respect to the score probabilities of *X* and *Y*, and $\begin{pmatrix} \mathbf{C_r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_s} \end{pmatrix}$ is made up of the two *C*-matrices, $\mathbf{C_r}$ and $\mathbf{C_s}$, computed in Step 1. In (4), $\|x\| = \sqrt{\sum_j \boldsymbol{x}_j^2}$ denotes the Euclidian length (norm) of vector ***x***.

Population Invariance Measures Extension of Step 5

Significance test for two subgroups' equating functions. When loglinear smoothing models and equating functions are computed for independent subpopulations, the differences between their equated scores can be evaluated with respect to a standard error of equating difference (SEED). The SEED used in this paper differs from what was proposed in von Davier et al. (2004) because the equated score differences evaluated in this paper are of independent subpopulations (where the *C*-matrices are not shared) rather than of linear and curvilinear equating functions (where the *C*-matrices are common). The SEED used in this paper is the square root of the sum of each subpopulation's squared standard errors of equating (SEE; see Appendix A).

$$SEED_Y(x) == \sqrt{\left\| \left(\frac{\partial e_{YP1}}{\partial \boldsymbol{r_{P1}}}, \frac{\partial e_{YP1}}{\partial \boldsymbol{s_{P1}}} \right) \begin{pmatrix} \mathbf{C_{rP1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_{sP1}} \end{pmatrix} \right\|^2 + \left\| \left(\frac{\partial e_{YP2}}{\partial \boldsymbol{r_{P2}}}, \frac{\partial e_{YP2}}{\partial \boldsymbol{s_{P2}}} \right) \begin{pmatrix} \mathbf{C_{rP2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_{sP2}} \end{pmatrix} \right\|^2} \tag{5}$$

$$= \sqrt{Var(\hat{e}_{YP1}) + Var(\hat{e}_{YP2})}$$

Thus, the SEED is not unique to the kernel method, but can be, and has been (Harris & Kolen, 1986, p. 40), computed based on the standard errors of any equating function by noting that the standard error of the difference of the equating functions of two independent subpopulations is the square root of the sum of each equating function's variance evaluated at each score of *X*.

The standard error of the RMSD, RMSDSE. The RMSD measure gives a value based on a particular *X* score that is a function of the estimated probabilities in the subpopulations and population ($RMSD(x; \hat{\boldsymbol{r}}_{\boldsymbol{P1}}, \hat{\boldsymbol{r}}_{\boldsymbol{P2}}, ..., \hat{\boldsymbol{r}}_{\boldsymbol{PG}}; \hat{\boldsymbol{s}}_{\boldsymbol{P1}}, \hat{\boldsymbol{s}}_{\boldsymbol{P2}}, ..., \hat{\boldsymbol{s}}_{\boldsymbol{PG}})$). The formula for the standard error for the RMSD (RMSDSE) can be written in a form that is general enough to apply to all of the major equating designs. This formula includes the derivatives of the RMSD and subpopulation equating functions with respect to all P_g *X* and *Y* score probabilities, derivatives of all P_g *X* and *Y* score probabilities with respect to the estimated (smoothed) distributions, and the factorized variance-covariance matrices of all P_g estimated distributions:

$RMSDSE(x) =$

$$\left\|\left[\frac{\partial RMSD}{\partial \boldsymbol{r}_{P1}},....,\frac{\partial RMSD}{\partial \boldsymbol{r}_{PG}},\frac{\partial RMSD}{\partial \boldsymbol{s}_{P1}},...,\frac{\partial RMSD}{\partial \boldsymbol{s}_{PG}}\right]\left[\mathbf{J}_{\mathbf{DF}}\mathbf{C}\right]\right\| =$$

$$\left\|\left[\frac{\partial RMSD}{\partial \boldsymbol{r}_{P1}},....,\frac{\partial RMSD}{\partial \boldsymbol{r}_{PG}},\frac{\partial RMSD}{\partial \boldsymbol{s}_{P1}},...,\frac{\partial RMSD}{\partial \boldsymbol{s}_{PG}}\right]\begin{bmatrix} \begin{pmatrix} \mathbf{C}_{\mathbf{rP1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{rP2}} & \mathbf{0} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{C}_{\mathbf{rPG}} \end{pmatrix} & \begin{pmatrix} \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \end{pmatrix} \\ \begin{pmatrix} \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{0} \end{pmatrix} & \begin{pmatrix} \mathbf{C}_{\mathbf{sP1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{sP2}} & \mathbf{0} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \mathbf{0} & \mathbf{0}\ldots\ldots & \mathbf{C}_{\mathbf{sPG}} \end{pmatrix} \end{bmatrix}\right\| \quad (6)$$

The derivatives of RMSD with respect to the P_g *X* and *Y* score probabilities are given in Appendix B. An accompanying paper will capitalize on the generality provided by von Davier et al.'s (2004) development of $\mathbf{J}_{\mathbf{DF}}\mathbf{C}_{\mathbf{Pg}}$, and show how the RMSD standard errors can be computed in population invariance studies using data collection designs other than the equivalent groups design.

Example

In this section, the previously described population invariance measures and standard errors are demonstrated and evaluated using actual test data. The data were obtained in a special study where two 42-item exam forms were given to high school students in a spiraled administration. The content of the exam was English literature. These data were used to assess the extent of population invariance in the equating function for the two exam forms. The subpopulations of interest were examinees from schools that were not in large cities (*P1*) and examinees from schools that were in large cities (*P2*). Table 3 presents the summary statistics for the population and subpopulations based on number-correct scores. The statistics in Table 3 show that the large sample of *P1* examinees did better on the *X* and *Y* forms than did the smaller sample of *P2* examinees. In addition, the *P1* examinees did slightly better on the *X* form than on the *Y* form while the *P2* examinees did slightly better on the *Y* form than on the *X* form.

Table 3

Descriptive Statistics of the Subpopulations and Population

Subpopulations and population	Test	N	Mean	Std. dev.	Skew	Kurtosis
$P1$	X	973	24.89	6.65	.07	-.47
$P1$	Y	958	24.85	6.45	.02	-.21
$P2$	X	296	23.87	7.39	-.15	-.45
$P2$	Y	294	23.92	7.25	-.07	-.54
P (= $P1$ + $P2$)	X	1,269	24.65	6.84	.01	-.35
P (= $P1$ + $P2$)	Y	1,252	24.63	6.66	-.03	-.17

Kernel Equating, Steps 1-4

Loglinear smoothing models were fit to the four subpopulations' score distributions. The model selection process was to consider the relative fits of five models that preserved 2 through 6 moments using four likelihood ratio chi-square tests with alpha levels of $1\text{-}(1\text{-}.05)^{1/4} = .0127$ (Haberman, 1974). The selected models for *X* and *Y* in *P1* both preserved 6 and 5 moments and had likelihood ratio statistics of 17.11 ($df = 36$, $p > .05$) and 27.58 ($df = 37$, $p > .05$), respectively. The selected models for *X* and *Y* in *P2* preserved 2 moments and had likelihood ratio statistics of 26.16 ($df = 40$, $p > .05$) and 49.84 ($df = 40$, $p > .05$), respectively. The score probabilities were obtained directly from the smoothed frequencies.

Continuous cdfs were estimated based on the discrete and smoothed score distributions of the four subpopulation score distributions and two population (*P1 + P2)* score distributions. A parabolic interpolation procedure (Press, Teukolsky, Vetterling, & Flannery, 1992) was used to select Gaussian kernel bandwidths that minimized the extent to which the continuous distributions deviated from the loglinear smoothed distributions, while having very few modes.

Finally, the *X*-to-*Y* kernel equating functions for *P1*, *P2*, and *P* were computed, along with the RMSD. Figure 1 plots the equated score differences (*P1 - P2*), along with practical differences-that-matter lines of +/- .5 score points. Figure 2 plots the RMSD, along with a practical difference-that-matters line of $.5/\sigma_{YP}$. The figures indicate that population dependence in the X-to-Y equating function is serious enough to create practically important differences between the subpopulations' equating functions (Figure 1) and between each of the subpopulation's and the population equating functions (Figure 2) for scores $X = 0$ through 9.

Step 5: Standard Errors, Delta Estimates

The population invariance measures were evaluated with respect to statistical significance. Figures 3 and 4 plot the equated score differences and the RMSD values, along with the differences-that-matter lines and +/- 2 times the delta method standard errors. Figures 3 and 4 indicate that the equated score differences and RMSD values are statistically significant for scores $X = 0$ through 6. Both figures indicate that the population invariance measures become extremely variable around score $X = 7$, a score point where the data are unusually sparse relative to the frequencies suggested by the loglinear smoothing models. The estimated variability of the RMSD values is directly determined by the magnitude of the equated score differences (see Appendix B), which accounts for the abrupt decreases in its standard error at scores $X = 19$ and 22, the X scores with the smallest equated score differences.

Evaluating the Accuracy of the Delta Standard Errors

The delta method standard error estimates were evaluated with respect to empirical variability. Two hundred datasets for three sample size conditions were simulated from each of the loglinear smoothing models selected for the four univariate distributions. For one sample size condition, the four sample sizes of the original data were used ($N_{XP1} = 973$, $N_{YP1} = 958$, $N_{XP2} = 296$, $N_{YP2} = 294$). For the other two sample size conditions, the four distributions were generated with equal sample sizes of 300 and 1,000. The kernel equating functions were then computed for each of these datasets and the same smoothing models and bandwidth parameters that were selected with the actual data. Averages of the 200 delta method standard errors were then computed and evaluated with respect to the standard deviations of the population invariance measures. This evaluation provided an estimate of empirical variability while adhering to the assumptions of the delta method (i.e., the assumptions that the loglinear models are the true models and that the same smoothing models and bandwidth parameters are used across all replications).

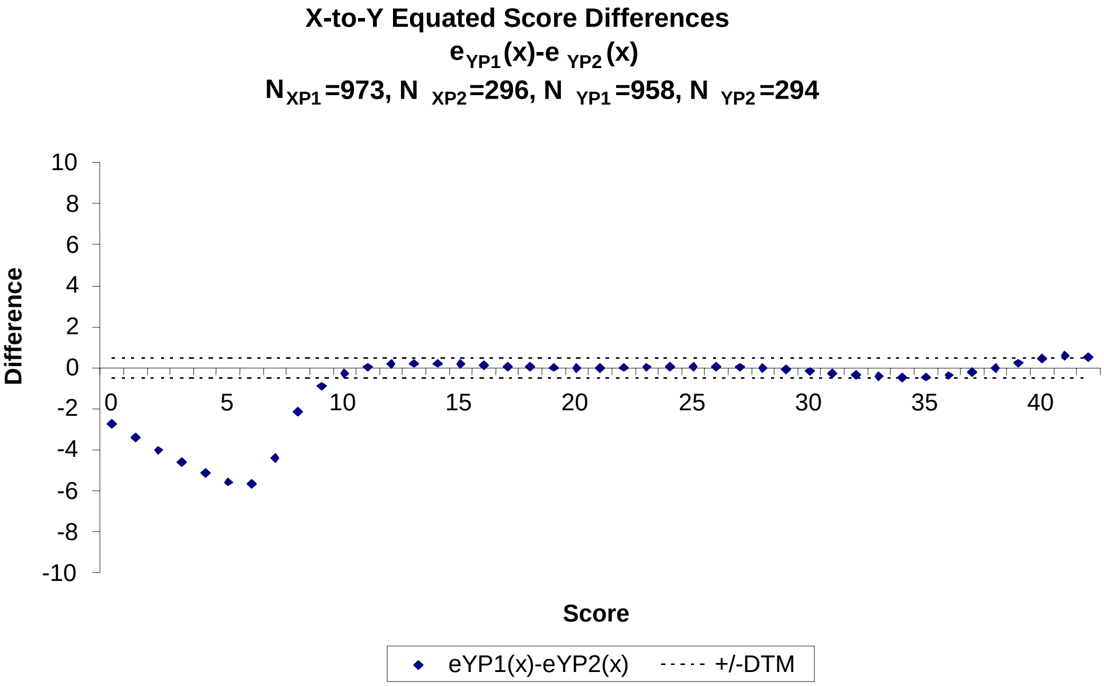

Figure 1. X-to-*Y* equated score differences.

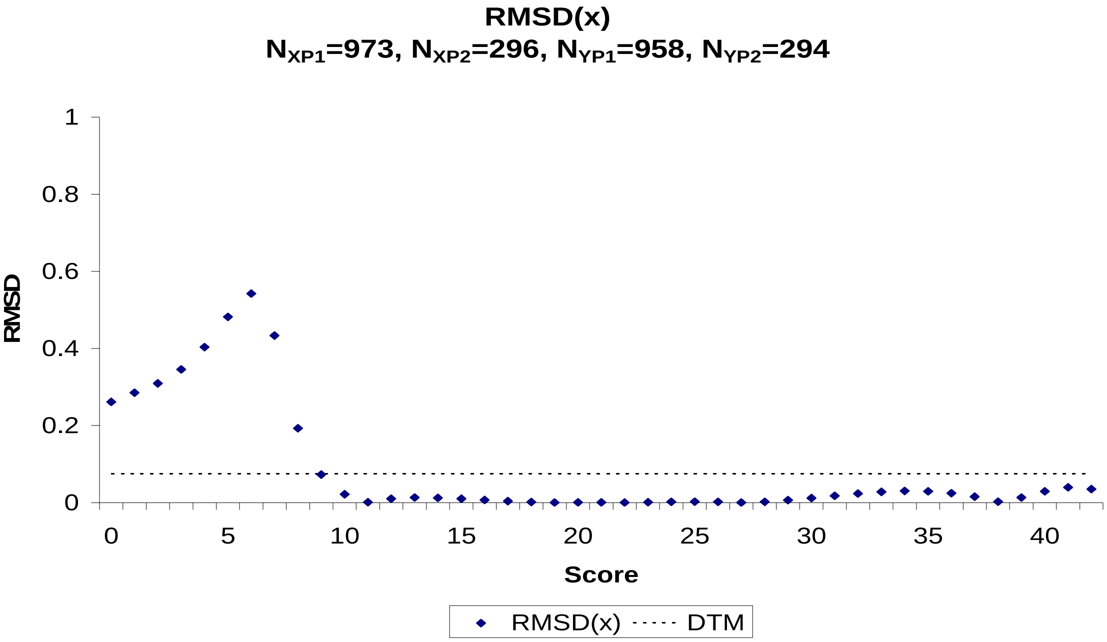

Figure 2. RMSD(x).

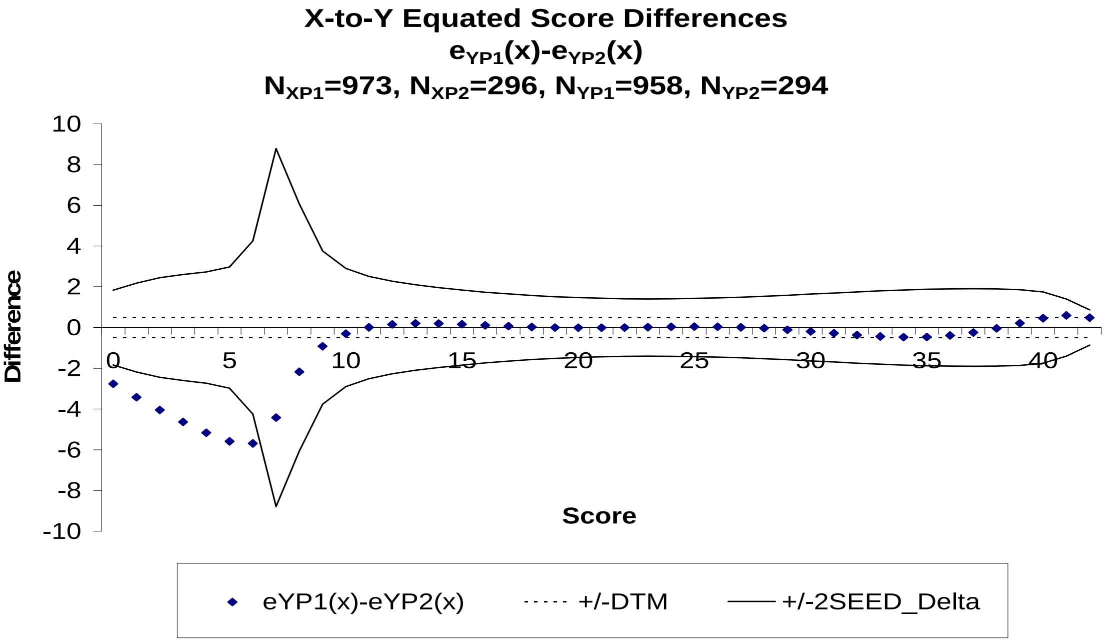

Figure 3. ***X*-to-*Y* equated score differences.**

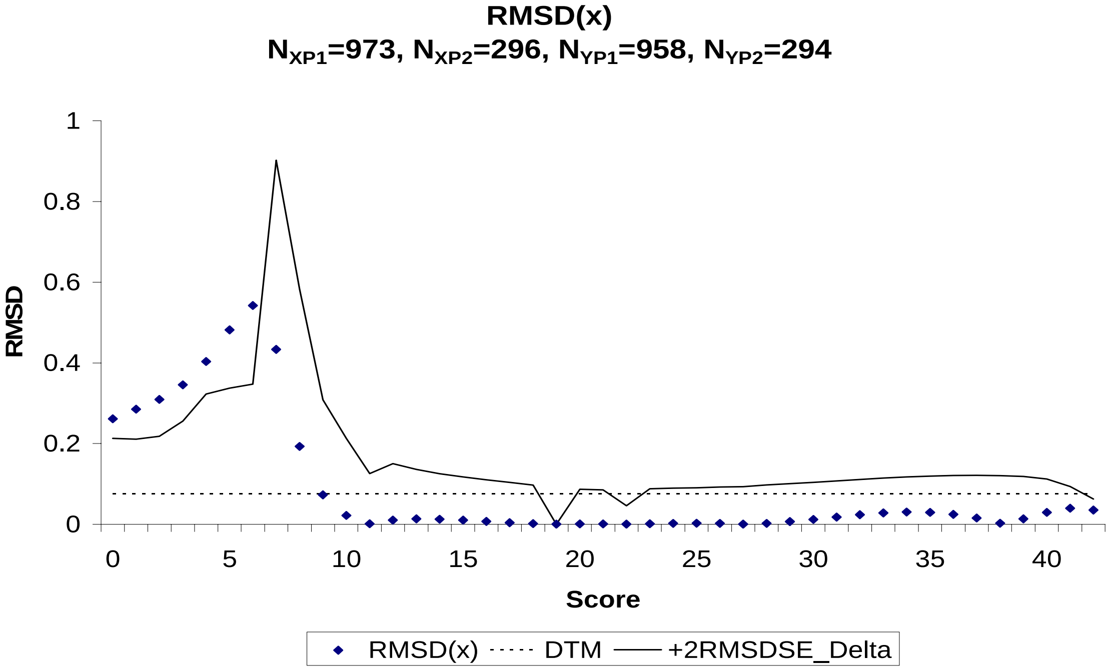

Figure 4. **RMSD(*x*).**

Figures 5, 6, and 7 plot the equated score differences and +/- two times the average of the delta method standard error estimates and +/- two times the empirical standard deviations. These three figures show fairly close agreement between the delta method standard error estimates and the empirical standard deviations. The delta method estimates are closest to the empirical standard deviations for the sample size condition of N = 1,000 for all four distributions (Figure 7). In Figure 5 (N_{XP1} = 973, N_{YP1} = 958, N_{XP2} = 296, N_{YP2} = 294) and Figure 7 (N = 1,000), the equated score differences at X = 0 through 6 are statistically significant based on the average delta method standard errors and also on the empirical standard deviations. In Figure 6, (N = 300), none of the equated score differences are statistically significant based on the average delta method standard errors and on the empirical standard deviations.

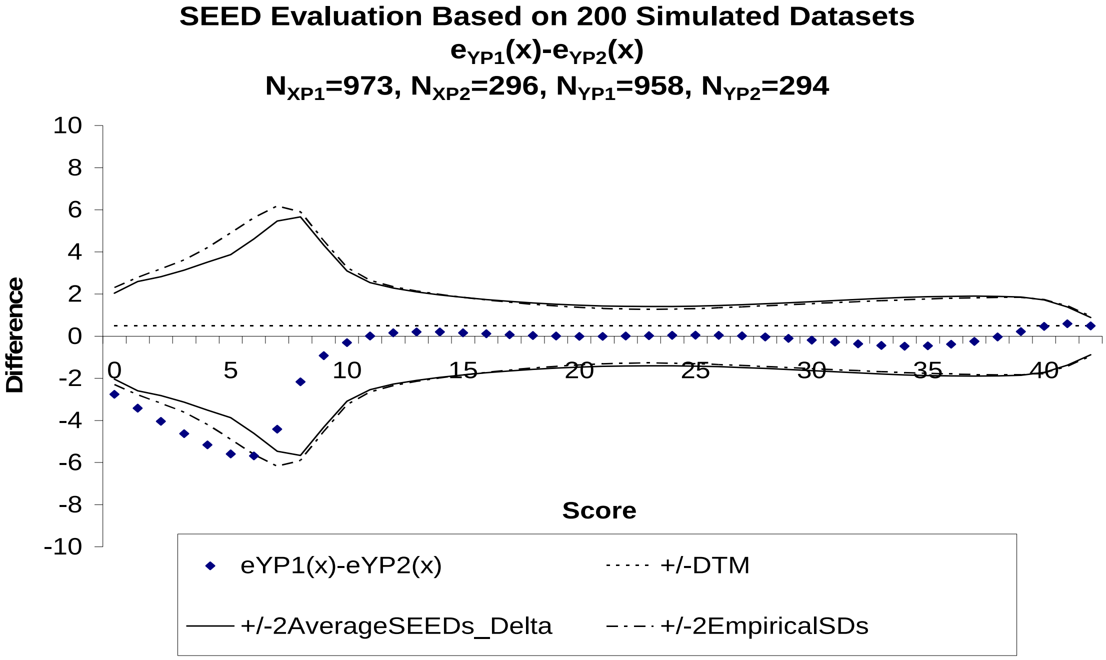

Figure 5. **SEED evaluation based on 200 simulated datasets.**

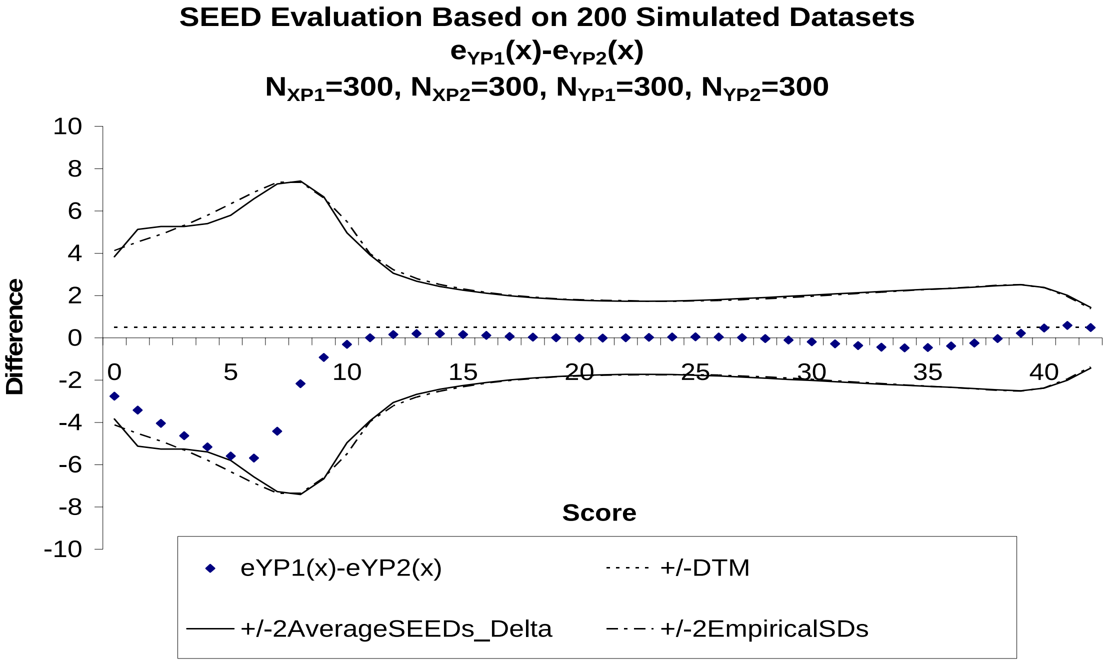

Figure 6. **SEED evaluation based on 200 simulated datasets.**

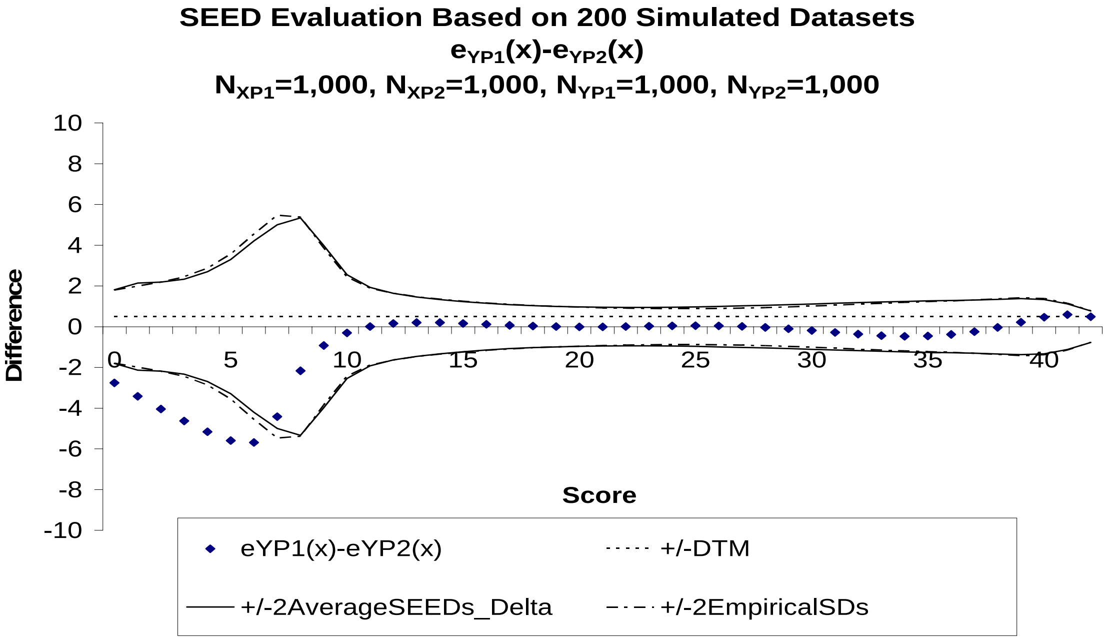

Figure 7. **SEED evaluation based on 200 simulated datasets.**

Figures 8, 9, and 10 evaluate the variability estimates of the RMSD. The RMSD depends on the differences in populations' sample sizes (3), so the reader should note that the RMSD values based on the sample sizes in the observed data (Figures 2, 4, and 8) differ from the RMSD values based on equal sample sizes in all four distributions (Figures 9 and 10). Figures 8, 9, and 10 show that the average delta method standard errors consistently overestimate the RMSD's empirical variability. The differences between the delta method's standard errors and the empirical standard deviations are greatest when the four univariate distributions are based on sample sizes of 300 (Figure 9). For Figure 9, none of the RMSD values are statistically significant based on the average delta method standard errors, while the RMSD values at scores X = 3 through 5 are statistically significant (but barely) based on the empirical standard deviations. For Figure 8 (N_{XP1} = 973, N_{YP1} = 958, N_{XP2} = 296, N_{YP2} = 294) and Figure 10 (N = 1,000), the delta method standard errors are sufficiently close to the empirical standard deviations so that they agree on which RMSD values are statistically significant (X = 0 through 6) and insignificant (X = 7 through 42).

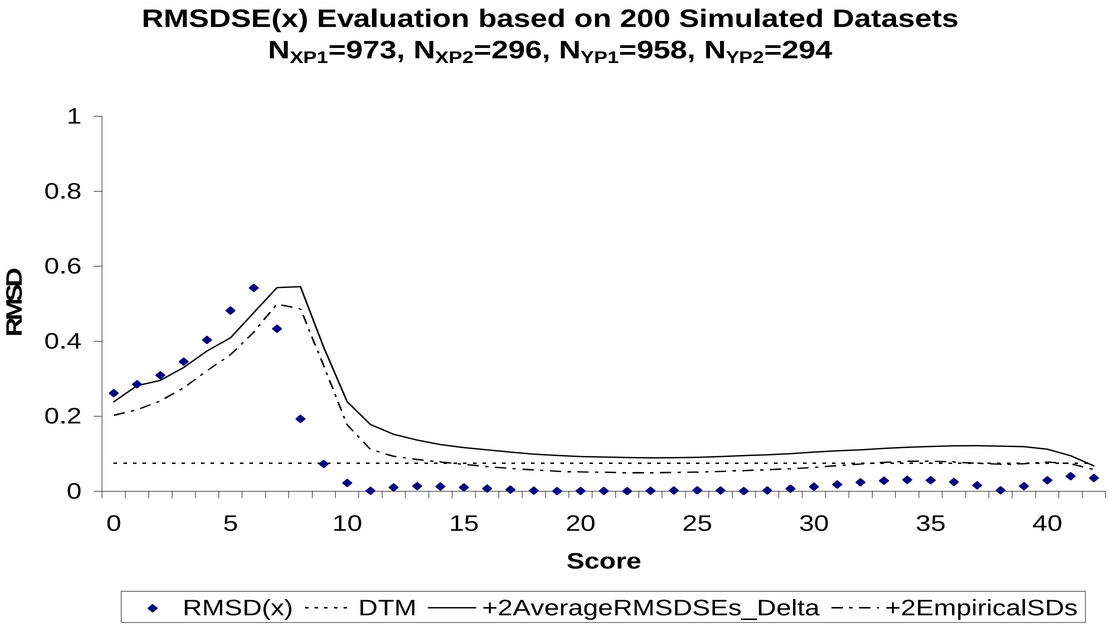

***Figure 8.* RMSDSE(*x*) evaluation based on 200 simulated datasets.**

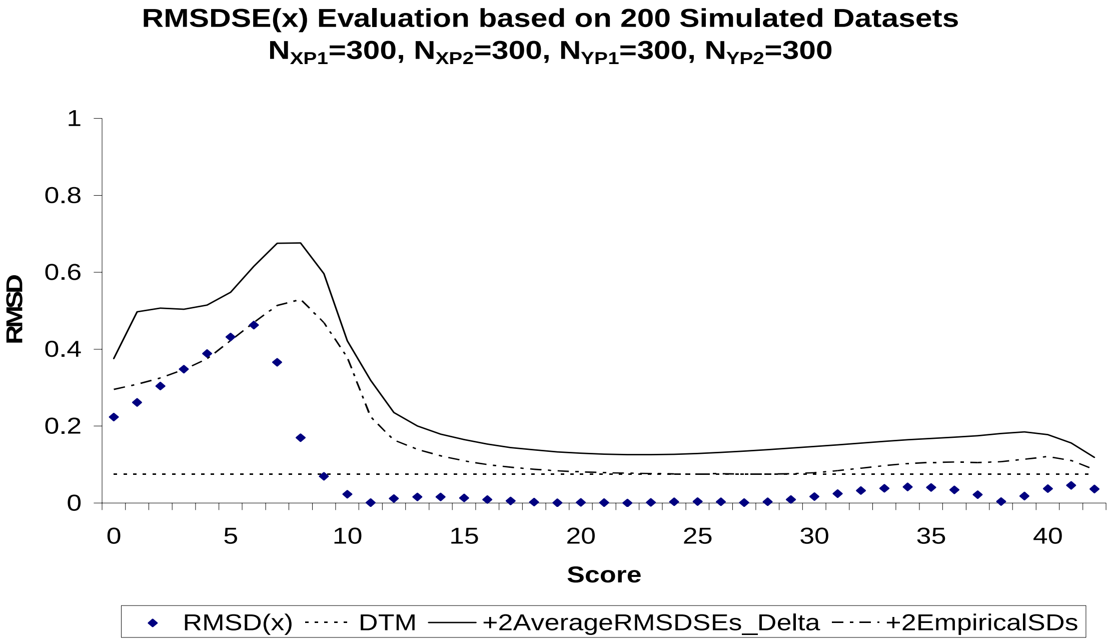

Figure 9. **RMSDSE(*x*) evaluation based on 200 simulated datasets.**

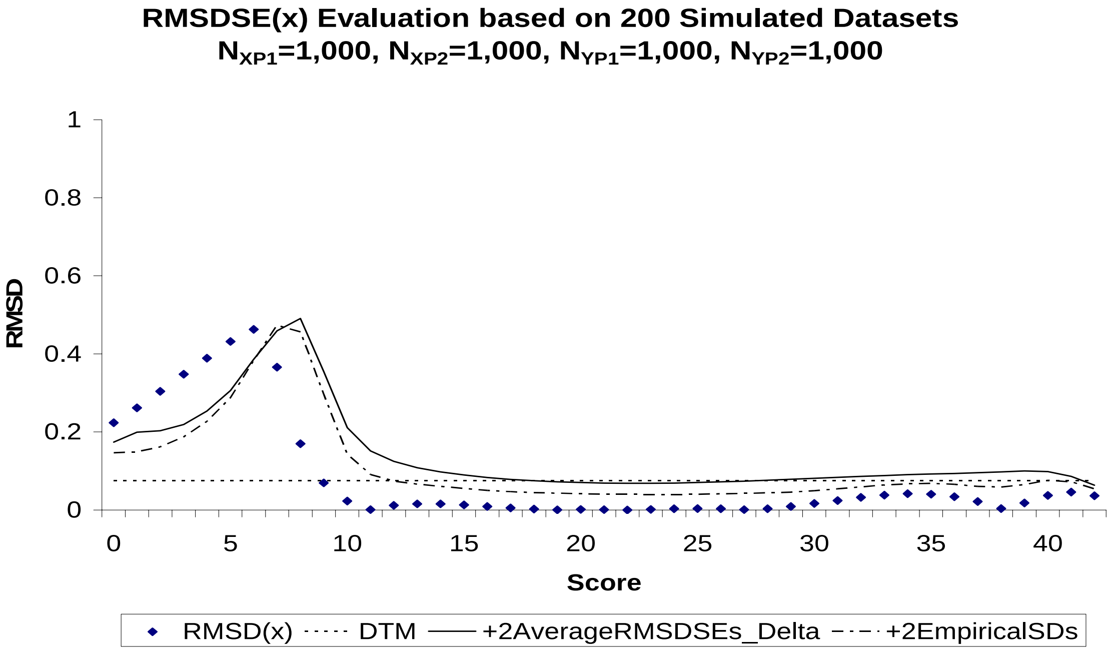

Figure 10. **RMSDSE(*x*) evaluation based on 200 simulated datasets.**

Discussion

This paper demonstrates how equating functions’ sampling variabilities can be estimated and incorporated into evaluations of population invariance. The kernel method and its delta method standard errors were extended to compute the standard errors of differences between subpopulations’ equating functions and standard errors of the RMSD. These standard errors were demonstrated on actual test data and evaluated in terms of equating function variability from data that were simulated from the same conditions as the actual data. The delta method standard error estimates of equated score differences more closely approximated actual sampling variability than did the delta method standard error estimates of the RMSD. The delta method estimates for the equated score differences and the RMSD were closer to actual variability when the equating functions were computed based on large, rather than small, sample sizes, which was expected from the literature (Jarjoura & Kolen, 1985; Liou & Cheng, 1995; Liou, Cheng, & Johnson, 1997). The delta estimates would not have been expected to work as well had they been evaluated with respect to other equating decisions, such as the selection of the appropriate loglinear model and/or bandwidth parameter. These additional decisions are not incorporated in the delta method estimates, but are certainly relevant to the decisions made in practice that add variability to all aspects of equating.

Many extensions of this work are possible. Because the derivations given in this paper utilize the kernel method’s general framework, the computation of the standard errors for population invariance measures in all of the major equating designs are straightforward. Population invariance evaluations in the single group, counterbalanced, and non-equivalent groups with anchor test designs are more complex than in the equivalent groups design because they are based on bivariate frequency tables of highly correlated tests that are usually sparse and require more complicated loglinear models and model search strategies. In addition to extending this work to other equating designs, wide ranges of sample size, degrees of true and false population invariance, and situations with more than two subpopulations could also be considered. These extensions would be informative for investigations of population invariance that commonly evaluate population invariance with respect to practical, rather than statistical, significance.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*(4), 327–345.

Bishop, Y. M. M., Feinberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

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

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, M. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS RM-94-10). Princeton, NJ: ETS.

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

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2004). Invariance of score linking across gender groups for three Advanced Placement Program examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR-03-27, pp. 79–118). Princeton, NJ: ETS.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30,* 589–600.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35–43.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (ETS Technical Report 89-83). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Technical Report 89-84). Princeton, NJ: ETS.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10,* 143–160.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). New York: Macmillan.

Kolen, M. J., & Brennan, R. J. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349–69.

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20,* 259–286.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. pp 221–262). New York: Macmillan.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). Washington, DC: American Psychological Association.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement, 35*(4), 277–296.

Yang, W. L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*(1), 33–41.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004, July). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*(4), 274–289.

Appendix A

The Standard Error of Equating Difference for Independent Subgroups' Equating Functions (Equivalent Groups Design)

Let the two subgroups be *P1* and *P2*.

Let the row vector of partial derivatives of the equating function with respect to the score probabilities of *X* and *Y* for each subpopulation be: $\mathbf{J}_{eY,P1} = \left(\frac{\partial e_{YP1}}{\partial \boldsymbol{r}_{\boldsymbol{P1}}}, \frac{\partial e_{YP1}}{\partial \boldsymbol{s}_{\boldsymbol{P1}}} \right)$ and $\mathbf{J}_{eY,P2} = \left(\frac{\partial e_{YP2}}{\partial \boldsymbol{r}_{\boldsymbol{P2}}}, \frac{\partial e_{YP2}}{\partial \boldsymbol{s}_{\boldsymbol{P2}}} \right)$.

For the equivalent groups design, the product of the design function and *C*-matrices for each subpopulation is $\mathbf{J_{DF}C_{P1}} = \begin{pmatrix} \mathbf{C_{rP1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_{sP1}} \end{pmatrix}$ and $\mathbf{J_{DF}C_{P2}} = \begin{pmatrix} \mathbf{C_{rP2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_{sP2}} \end{pmatrix}$. Under the assumption of subgroup independence, create $\mathbf{J_{DF}C_{P1\&2}}$ as $\begin{pmatrix} \mathbf{J_{DF}C_{P1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_{DF}C_{P2}} \end{pmatrix}$.

Then based on the delta method,

$$Var(\widehat{e}_{YP1} - \widehat{e}_{YP2}) = \left(\mathbf{J}_{eYP1}, -\mathbf{J}_{eYP2} \right) \begin{pmatrix} \mathbf{J_{DF}C_{P1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_{DF}C_{P2}} \end{pmatrix} \begin{pmatrix} \left[\mathbf{J_{DF}C_{P1}}\right]^{\mathrm{t}} & \mathbf{0} \\ \mathbf{0} & \left[\mathbf{J_{DF}C_{P2}}\right]^{\mathrm{t}} \end{pmatrix} \begin{pmatrix} \mathbf{J}_{eYP1}{}^{t} \\ -\mathbf{J}_{eYP2}{}^{t} \end{pmatrix}$$

$$= \left(\mathbf{J}_{eYP1}, -\mathbf{J}_{eYP2} \right) \begin{pmatrix} \mathbf{J_{DF}C_{P1}}\left[\mathbf{J_{DF}C_{P1}}\right]^{\mathrm{t}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J_{DF}C_{P2}}\left[\mathbf{J_{DF}C_{P2}}\right]^{\mathrm{t}} \end{pmatrix} \begin{pmatrix} \mathbf{J}_{eYP1}{}^{t} \\ -\mathbf{J}_{eYP2}{}^{t} \end{pmatrix}$$

$$= \left(\mathbf{J}_{eYP1}\mathbf{J_{DF}C_{P1}}\left[\mathbf{J_{DF}C_{P1}}\right]^{\mathrm{t}}, -\mathbf{J}_{eYP2}\mathbf{J_{DF}C_{P2}}\left[\mathbf{J_{DF}C_{P2}}\right]^{t} \right) \begin{pmatrix} \mathbf{J}_{eYP1}{}^{t} \\ -\mathbf{J}_{eYP2}{}^{t} \end{pmatrix}$$

$$= \left(\mathbf{J}_{eYP1}\mathbf{J_{DF}C_{P1}}\left[\mathbf{J_{DF}C_{P1}}\right]^{\mathrm{t}} \mathbf{J}_{eYP1}{}^{t} + \mathbf{J}_{eYP2}\mathbf{J_{DF}C_{P2}}\left[\mathbf{J_{DF}C_{P2}}\right]^{t} \mathbf{J}_{eYP2}{}^{t} \right)$$

$$= \left\| \mathbf{J}_{eYP1}\mathbf{J_{DF}C_{P1}} \right\|^{2} + \left\| \mathbf{J}_{eYP2}\mathbf{J_{DF}C_{P2}} \right\|^{2}$$

$$= Var(\widehat{e}_{YP1}) + Var(\widehat{e}_{YP2})$$

Therefore, the SEED between independent subpopulations' equating functions is $\sqrt{Var(\widehat{e}_{YP1}) + Var(\widehat{e}_{YP2})}$.

Appendix B

Derivatives of the RMSD with Respect to the Score Probabilities

The Derivative of RMSD With Respect to $\widehat{r}_{jPg}$.

$$RMSD(x_j) = \frac{\sqrt{\sum_g w_g [e_{YPg}(x_j) - e_{YP}(x_j)]^2}}{\sigma_{YP}}$$

By the chain rule,

$$\frac{\partial RMSD(x_j)}{\partial \widehat{r}_{jPg}} = \frac{\frac{\partial}{\partial \widehat{r}_{jPg}}\left(\sum_g w_g [e_{YPg}(x_j) - e_{YP}(x_j)]^2\right)}{2\sigma_{YP}\sqrt{\sum_g w_g [e_{YPg}(x_j) - e_{YP}(x_j)]^2}} .$$

By the chain rule,

$$\left[\frac{\partial e_{YP}(x_j)}{\partial \widehat{r}_{jPg}} = \frac{\partial e_{YP}(x_j)}{\partial \widehat{r}_{jP}} \frac{\partial \widehat{r}_{jP}}{\partial \widehat{r}_{jPg}} = \frac{\partial e_{YP}(x_j)}{\partial \widehat{r}_{jP}} \frac{n_{XPg}}{\sum_g^G n_{XPg}}\right],$$

$$\frac{\partial RMSD(x_j)}{\partial \hat{r}_{jPg}} = \frac{w_g[e_{YPg}(x_j) - e_{YP}(x_j)]\left(\frac{\partial e_{YPg}(x_j)}{\partial \hat{r}_{jPg}} - \frac{\partial e_{YP}(x_j)}{\partial \hat{r}_{jP}}\frac{n_{XPg}}{\sum_{g}^{G} n_{XPg}}\right) + \sum_{h\neq g}^{G} w_h[e_{YPh}(x_j) - e_{YP}(x_j)]\left(-\frac{\partial e_{YP}(x_j)}{\partial \hat{r}_{jP}}\frac{n_{XPg}}{\sum_{g}^{G} n_{XPg}}\right)}{\sigma_{YP}\sqrt{\sum_g w_g[e_{YPg}(x_j) - e_{YP}(x_j)]^2}}.$$

Then $\frac{\partial e_{YPg}(x_j)}{\partial \hat{r}_{jPg}}$ and $\frac{\partial e_{YP}(x_j)}{\partial \hat{r}_{jP}}$ can be computed from equations given in Holland, King, and Thayer (1989) and in von Davier et al. (2004).

The Derivative of σ_{YP} With Respect to $\hat{s}_{kPg}$.

The derivative of RMSD with respect to $\hat{s}_{k,Pg}$ first requires the differentiation of σ_{YP} with respect to $\hat{s}_{k,Pg}$ and then an application of the quotient rule.

$$\sigma_{YP} = \sqrt{\sum_k (y_k - \mu_{YP})^2 \hat{s}_{kP}}$$

$$\frac{\partial \sigma_{YP}}{\partial \hat{s}_{kP}} = \frac{\frac{\partial}{\partial \hat{s}_{kP}}\left(\sum_k (y_k - \mu_{YP})^2 \hat{s}_{kP}\right)}{2\sqrt{\sum_k (y_k - \mu_{YP})^2 \hat{s}_{kP}}}.$$

From the appendix in Holland et al. (1989), $\frac{\partial \sigma_{YP}}{\partial \hat{s}_{kP}} = \frac{(y_k - \mu_{YP})^2}{2\sqrt{\sum_k (y_k - \mu_{YP})^2 \hat{s}_{kP}}}.$

To obtain the derivative with respect to $\hat{s}_{k,Pg}$, we apply the chain rule and multiply by the derivative of $\hat{s}_{kP}$ with respect to $\hat{s}_{kPg}$:

$$\begin{aligned}\frac{\partial\sigma_{YP}}{\partial\hat{s}_{kPg}} &= \frac{(y_k-\mu_{YP})^2}{2\sqrt{\sum_k (y_k-\mu_{YP})^2\,\hat{s}_{kP}}}\frac{n_{YPg}}{\sum_g^G n_{YPg}} \\ &= \frac{(y_k-\mu_{YP})^2\,n_{YPg}}{2\sigma_{YP}\sum_g^G n_{YPg}}\end{aligned}.$$

The Derivative of RMSD With Respect to $\hat{s}_{kPg}$.

$$RMSD(x_j)=\frac{\sqrt{\sum_g w_g\,[e_{YPg}(x_j)-e_{YP}(x_j)]^2}}{\sigma_{YP}}.$$

$$\frac{\partial RMSD(x_j)}{\partial\hat{s}_{kPg}}=\sigma_{YP}\,\frac{\frac{\partial}{\partial\hat{s}_{kPg}}\sum_g w_g\,[e_{YPg}(x_j)-e_{YP}(x_j)]^2}{2\sigma^2{}_{YP}\sqrt{\sum_g w_g\,[e_{YPg}(x_j)-e_{YP}(x_j)]^2}}-\frac{\sqrt{\sum_g w_g\,[e_{YPg}(x_j)-e_{YP}(x_j)]^2}\;\frac{\partial}{\partial\hat{s}_{kPg}}\left(\sigma_{YP}\right)}{\sigma^2{}_{YP}}$$

$$\frac{\partial RMSD(x_j)}{\partial \widehat{s}_{kPg}} =$$

$$\frac{w_g[e_{YPg}(x_j) - e_{YP}(x_j)]\left(\frac{\partial e_{YPg}(x_j)}{\partial \widehat{s}_{kPg}} - \frac{\partial e_{YP}(x_j)}{\partial \widehat{s}_{kP}} \frac{n_{YPg}}{\sum_g^G n_{YPg}} \right) + \sum_{h \neq g}^G w_h[e_{YPh}(x_j) - e_{YP}(x_j)]\left(- \frac{\partial e_{YP}(x_j)}{\partial \widehat{s}_{kP}} \frac{n_{YPg}}{\sum_g^G n_{YPg}} \right)}{\sigma_{YP}\sqrt{\sum_g w_g[e_{YPg}(x_j) - e_{YP}(x_j)]^2}} - \frac{RMSD(x_j)(y_k - \mu_{YP})^2 n_{YPg}}{2\sigma_{YP}^2 \sum_g^G n_{YPg}}$$

Then $\frac{\partial e_{YPg}(x_j)}{\partial \widehat{s}_{kPg}}$ and $\frac{\partial e_{YP}(x_j)}{\partial \widehat{s}_{kP}}$ can be computed from equations given in Holland et al. (1989) and in von Davier et al. (2004).