An Evaluation of Statistical Strategies for Making Equating Function Selections

Tim Moses

November 2008

ETS RR-08-60

An Evaluation of Statistical Strategies for Making Equating Function Selections

Tim Moses
ETS, Princeton, NJ

November 2008

Abstract

Nine statistical strategies for selecting equating functions in an equivalent groups design were evaluated. The strategies of interest were likelihood ratio chi-square tests, regression tests, Kolmogorov-Smirnov tests, and significance tests for equated score differences. The most accurate strategies in the study were the likelihood ratio tests and the significance tests for equated score differences.

Key words: Significance tests, equivalent groups equating functions

Acknowledgments

The author thanks Neil Dorans, Dan Eignor, Skip Livingston, and Frank Rijmen for helpful suggestions and Kim Fryer for editorial work.

Table of Contents

Page

Introduction ... 1

The Equivalent-Groups Equating Design and Possible Equating Functions ... 1

Statistical Significance Tests Proposed for Equating Function Selection ... 1

This Study ... 8

Method ... 8

Study Design ... 8

Results ... 14

Type I Error Results ... 14

Power Results ... 15

Score-Level Type I Error Rates of Test of Equated Score Differences ... 15

Score-Level Power Rates of Test of Equated Score Differences ... 18

Discussion ... 20

Summary ... 20

Issues With Each Significance Test ... 20

Suggestions for Practice ... 22

References ... 23

Notes ... 25

List of Tables

Page

Table 1. Descriptive Statistics for the Population Distributions Used in the Study 10

Table 2. Equating Function Selection Rates: Identity Population Equating Function............. 16

Table 3. Equating Function Selection Rates: Equipercentile Population Equating Function... 16

List of Figures

Page

Figure 1. Population distributions for the identity equating function condition....................... 11

Figure 2. Population distributions for the equipercentile equating function condition. 11

Figure 3. Equipercentile-linear, equipercentile-identity, and linear-identity equated score differences for the identity population equating function condition.......................... 12

Figure 4. Equipercentile-linear equated score differences for the equipercentile population equating function condition. 12

Figure 5. Equipercentile-identity equated score differences for the equipercentile population equating function condition. 13

Figure 6. Linear-identity equated score differences for the equipercentile population equating function condition. 13

Figure 7. Type I error rates for score-level tests of equipercentile-linear equated score differences (identity population equating function condition).................................. 17

Figure 8. Type I error rates for score-level tests of equipercentile-identity equated score differences (identity population equating function condition).................................. 17

Figure 9. Type I error rates for score-level tests of linear-identity equated score differences (identity population equating function condition). 17

Figure 10. Power rates for score-level tests of equipercentile-linear equated score differences (equipercentile population equating function condition). 19

Figure 11. Power rates for score-level tests of equipercentile-identity equated score differences (equipercentile population equating function condition). 19

Figure 12. Power rates for score-level tests of linear-identity equated score differences (equipercentile population equating function condition). 19

Introduction

There are many views and proposals in the equating literature about using statistical significance tests for selecting equating functions (Budescu, 1987; Dorans & Lawrence, 1990; Hanson, 1996; Jaeger, 1981; Kolen & Brennan, 2004; Moses, Yang, & Wilson, 2007; von Davier, Holland, & Thayer, 2004). These views range from encouragement for the potential of significance tests to formalize equating decisions, to criticism because significance tests do not address the practical implications of equating function differences, to disagreement about the criteria on which to base the significance tests. Several different statistical significance tests have been proposed for equating function selection, mostly by demonstrating a single test on one or two datasets rather than by comparing the long-run accuracies of several tests. The purpose of this study is to compare the accuracies of several proposed significance tests in different equating situations and to make recommendations for the most accurate tests for equating practice.

The Equivalent-Groups Equating Design and Possible Equating Functions

The equating situation considered in this study is one where test score data are gathered in an equivalent-groups design, where randomly and independently drawn samples from a common population of examinees are given one of two alternate forms of a test, either form X or form Y. An equating function is sought to map the observed X scores to the observed Y scores in such a way that unintended difficulty differences between the forms are eliminated. Three possible X-to-Y equating functions are (a) the identity function, which treats X and Y as if they are of equal difficulty; (b) the linear function, which adjusts X's scores for differences in the means and variances of X and Y and (c) the equipercentile function, which finds Y scores with percentiles that equal those of each X score.

Statistical Significance Tests Proposed for Equating Function Selection

The statistical significance tests reviewed in this section have been proposed for comparing and selecting the identity, linear, and equipercentile equating functions in the context of the equivalent-groups equating design. All of the tests are based on comparing how well two models fit the observed test data, one model that corresponds to a relatively simple equating function (i.e., the null hypothesis) and another model that corresponds to a relatively complex equating function (i.e., the alternative hypothesis). The criteria for rejecting the null hypothesis and accepting the alternative hypothesis can be manipulated to give a determined amount of preference for the more complex model. This preference for the more complex model is traditionally defined so that incorrect rejections of the null hypotheses (i.e., Type I errors) are low (a rate usually defined at 0.05).

Likelihood ratio chi-square tests of loglinear models. One class of significance tests for equating functions is based on evaluating the differences in the *X* and *Y* frequency distributions using likelihood ratio chi-square tests of loglinear models (Hanson, 1996; Moses et al., 2007). The loglinear models relate fitted frequencies to polynomial functions of the test scores, where the polynomial terms in the model preserve moments (mean, variance, skew, etc.) of the observed distribution in the fitted distribution (Holland & Thayer, 1987, 2000). For example, the loglinear model of one test form's frequency distribution can be expressed as:

$$\log_e(m_j) = \beta_0 + \sum_{d=1}^{D} \beta_{(d+1)} j^d \ , \qquad (1)$$

where m_j is the fitted frequency at score *j*, β_0 is a constant that ensures that the fitted frequencies sum to the total sample size ($= \sum_{j=1}^{J} m_j = N$), the remaining β terms are parameters estimated by maximum likelihood estimation, and *D* determines the number of moments in the observed distribution that are preserved in the fitted distribution.

The use of loglinear models for comparing equating functions involves formulating models that allow for degrees of difference between the *X* and *Y* frequency distributions. A reduced loglinear model that corresponds to the identity equating function constrains all of the moments of the *X* and *Y* frequency distributions to be equal:

$$\text{Reduced model: } \log_e(m_{jk}) = \beta_0 + \beta_1 I(k) + \sum_{d=1}^{D} \beta_{(d+1)} j^d \ , \qquad (2)$$

where m_{jk} is the fitted frequency at score *j* of test form *k* (= *X* or *Y*), and $I(k)$ is an indicator function that defines when the m_{jk} *th* frequency is from *X* or *Y*.

An MSD (i.e., mean and standard deviation) loglinear model corresponds to the linear equating function and allows for differences between only the first (the mean) and second (the variance) moments of the *X* and *Y* frequency distributions (Angoff, 1971; von Davier, Holland,

& Thayer, 2004). This allowance of mean and variance differences is achieved by adding two product terms to the reduced model,

$$\text{MSD model: } \log_e(m_{jk}) = \beta_0 + \beta_1 I(k) + \sum_{d=1}^{D} \beta_{(d+1)} j^d + \sum_{d=1}^{2} \beta_{(D+d+1)} I(k) j^d \ . \qquad (3)$$

A full loglinear model that corresponds to the equipercentile equating function allows for differences between all of the fitted moments of the *X* and *Y* frequency distributions:

$$\text{Full model: } \log_e(m_{jk}) = \beta_0 + \beta_1 I(k) + \sum_{d=1}^{D} \beta_{(d+1)} j^d + \sum_{d=1}^{D} \beta_{(D+d+1)} I(k) j^d \ . \qquad (4)$$

The choice of the most appropriate loglinear model and its corresponding equating function involves using a series of statistical significance tests that compare how well different models fit the observed frequencies, n_{jk}. Each model's fit to the observed frequencies is summarized in a likelihood ratio chi-square statistic:

$$G^2 = 2\sum_k \sum_j n_{jk} \log_e(\frac{n_{jk}}{m_{jk}}) . \qquad (5)$$

The difference between two models' likelihood ratio chi-square statistics is tested for statistical significance by determining if the probability of this difference on a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated in the two models is less than a defined Type I error rate. This statistical significance test is known as the *likelihood ratio test*. A nonsignificant likelihood ratio test is statistical evidence that the parameters in the more complex model, and not in the simpler model, are not in the population model (i.e., there is support for the null hypothesis that the simpler model is correct). A significant likelihood ratio test is statistical evidence that the parameters in the more complex model are in the population model (i.e., there is support for the alternative hypothesis that the complex model is correct).

The first required decision for the appropriate loglinear model choice is to determine *D*, a choice that can be made by fitting full models with a range of *D* values (e.g., 3, 4, 5, … , 10), computing models' likelihood ratio chi-square statistics, and conducting likelihood ratio tests of the differences between likelihood ratio chi-square statistics for models with large *D*s and models with *D*s that are one-term simpler (e.g., $D = 10$ vs. $D = 9$, $D = 9$ vs. $D = 8$, … , $D = 4$ vs. $D = 3$). The chosen model is the one with the largest *D*, *D'*, that has a likelihood ratio chi-square that is significantly different from that of the model with $D = D' - 1$ (Hanson, 1996). A procedure from Haberman (1974) can be used to control the Type I error rate for all of the significance tests, α , by defining the Type I error rates for each individual test as $1-(1-\alpha)^{\frac{1}{P-1}}$, where *P* is the total number of considered models.

When *D* is determined, the reduced, MSD, and full models can be fit, their likelihood ratio chi-square statistics calculated, and the differences between two models' likelihood ratio chi-square statistics can be tested for statistical significance. The likelihood ratio test of the difference between the reduced and MSD models' likelihood ratio chi-square statistics is a comparison of the identity and linear equating functions. The likelihood ratio test of the difference between the reduced and full models' likelihood ratio chi-square statistics is a comparison of the identity and equipercentile functions. The likelihood ratio test of the difference between the MSD and full loglinear models' likelihood ratio chi-square statistics is a comparison of the linear and equipercentile functions. Like the significance tests used to select *D*, the significance tests for comparing the reduced, MSD, and full loglinear models can be arranged hierarchically (Hanson, 1996; Moses et al., 2007), proceeding by evaluating the most complex model/equating function (i.e., full/equipercentile) to simpler models/equating functions (i.e., MSD/linear and reduced/identity).

Kolmogorov-Smirnov distances. Another statistical significance test that has been proposed for equating function selection is based on assessing test forms' distribution differences through comparing their cumulative densities (Jaeger, 1981). Cumulative densities for *X* and *Y* can be computed from each of the *J* score observed frequencies, n_{jX} and n_{jY} , as:

$$F(x=j)=\sum_{i=1}^{j}\frac{n_{iX}}{\sum_{j=1}^{J}n_{jX}}=\sum_{i=1}^{j}\frac{n_{iX}}{N_X}\text{, and} \tag{6}$$

$$F(y=j)=\sum_{i=1}^{j}\frac{n_{iY}}{\sum_{j=1}^{J}n_{jY}}=\sum_{i=1}^{j}\frac{n_{iY}}{N_Y}. \tag{7}$$

The Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948) can be used to compare the *X* and *Y* cumulative densities based on the maximum absolute difference from the *J* possible differences:

$$\text{KS} = \max\left|F(x = j) - F(y = j)\right| \text{ for all } J. \tag{8}$$

To assess KS for statistical significance, it is compared to a critical value that corresponds to some defined Type I error rate, and if KS exceeds that critical value, the null hypothesis that the test score distributions are equal is rejected. A well-known (Conover, 1980; Wilcox, 2005) critical value for determining if KS is statistically significant at a Type I error rate of 0.05 is $1.36\sqrt{\frac{N_X + N_Y}{N_X N_Y}}$. Alternative procedures that compute exact probabilities for a KS test statistic exist (Wilcox, 2005), but these are computationally unwieldy for the large sample sizes often used in equating (N_X & N_Y > 500).

The KS test has a direct correspondence with equating function comparisons. The traditional equipercentile (Kolen & Brennan, 2004) and kernel (von Davier et al., 2004) equating methods use computations based on percentile ranks and kernel-smoothed cumulative densities, both of which are similar to (6) and (7). When the KS test is based on raw frequencies, as in (8), this comparison of the *X* and *Y* test score distributions amounts to comparisons of the raw equipercentile and identity equating functions (Jaeger, 1981).

An additional interest of this study is in the application of constraints to the test scores' frequencies when using the KS test. For example, one proposal for constraining the differences of the test scores' distributions being compared was made by Budescu (1987), where the *X* and *Y* distributions being compared were restricted to being unimodal and were allowed to vary only in their first four moments. The constraints of interest in this study allow the test score distributions to differ only in their means and variances, but not in their higher moments (i.e., the expected frequencies from the MSD loglinear model in Equation 3, m_{iX}, and m_{iY}). Using the KS test to compare test score distributions differing only in their means and variances amounts to a test of the linear and identity equating functions:

$$\text{KS}_{\text{Linear}} = \max\left|F(x=j)-F(y=j)\right| = \max\left|\sum_{\boldsymbol{i}=1}^{\boldsymbol{j}}\frac{m_{iX}}{\boldsymbol{N_X}} - \sum_{\boldsymbol{i}=1}^{\boldsymbol{j}}\frac{m_{iY}}{\boldsymbol{N_Y}}\right| \text{ for all } J. \tag{9}$$

Equated score differences. A direct approach to using statistical significance tests for selecting equating functions is to compute the equating functions of interest, compute their differences, and then assess the differences with respect to their standard error. For example, the *X*-to-*Y* linear equating function [$\widehat{e}_{Y,L}(x_j)$],the standard error of equating [$SEE_{Y,L}(x_j)$], and the *X*-to-*Y* identity equating function (x_j) can be computed for all *J* possible scores and used to test the difference between the linear and identity equating functions at score x_j (Dorans & Lawrence, 1990):

$$\frac{\widehat{e}_{Y,L}(x_j)-x_j}{SEE_{Y,L}(x_j)}. \tag{10}$$

A significance test for the difference between the equipercentile and identity functions can be conducted by using the equipercentile equating function [$\widehat{e}_{Y,E}(x_j)$] and the standard error of equating [$SEE_{Y,E}(x_j)$]:

$$\frac{\widehat{e}_{Y,E}(x_j)-x_j}{SEE_{Y,E}(x_j)}. \tag{11}$$

A significance test for the difference between the equipercentile and linear functions can be conducted based on the standard error of equating difference between the equipercentile and linear equating functions [$SEED_{Y,E-L}(x_j)$, von Davier et al., 2004]:

$$\frac{\widehat{e}_{Y,E}(x_j)-\widehat{e}_{Y,L}(x_j)}{SEED_{Y,E-L}(x_j)}. \tag{12}$$

For a defined Type I error rate of 0.05, (10), (11), and (12) would be considered statistically significant when their absolute value exceeds 1.96.

One issue with the test statistics in (10), (11), and (12) is that they do not amount to overall tests of equating functions, rather to non-independent score-level significance tests. A way to use the score-level significance tests for overall equating decisions would be to manipulate the overall Type I error rate of the *J*–total significance tests, and one of the simplest and most general procedures for doing so is referred to as the *Bonferroni procedure* (Lomax, 2001). For a defined Type I error rate for the *J*–total significance tests, α , the Bonferroni procedure would require that at least one of the *J* test statistics for (10, (11), or (12) would have a probability on a *z* distribution[1] that was less than $\frac{\alpha}{J}$.

Regression test. A significance test proposed to assess an equipercentile equating function's curvilinearity is based on regression analysis (Jaeger, 1981). The *X*-to-*Y* equipercentile equating function, $\widehat{e}_{Y,E}$, is computed and fit to two possible least squares regression models of the raw *X* scores:

$$\widehat{e}_{Y,E}(x_l) = B_0 + B_1 x_l + \varepsilon_{l,P13} \text{ and} \tag{13}$$

$$\widehat{e}_{Y,E}(x_l) = B_0 + B_1 x_l + B_2 x_l^2 + B_3 x_l^3 + \varepsilon_{l,P14}, \tag{14}$$

for individuals $l = 1$ to N_X in the group of examinees that take form *X*. Model (13) fits a straight line to the equipercentile equating function, and model (14) fits a cubic function to the equipercentile function. The comparison of fit in models (13) and (14) is made based on the differences in the models' R-squares values,

$$\left(\frac{R^2_{P13} - R^2_{P14}}{2}\right) \Bigg/ \left(\frac{1 - R^2_{P14}}{N_X - 4}\right), \tag{15}$$

where $R^2_{P13} = 1 - \frac{\sum_l \left(\varepsilon_{l,P13}\right)^2}{\sum_l \left(\widehat{e}_{Y,E}(x_l)\right)^2}$, $R^2_{P14} = 1 - \frac{\sum_l \left(\varepsilon_{l,P14}\right)^2}{\sum_l \left(\widehat{e}_{Y,E}(x_l)\right)^2}$), and (15) is evaluated by computing its probability on an F distribution with 2 and $N_X - 4$ degrees of freedom. When the probability of (15) is less than a defined Type I error level (e.g., less than 0.05), there is support for the alternative hypothesis of an equipercentile equating function with curvilinearity rather than the null hypothesis of an equipercentile equating function that is essentially linear (i.e., a linear equating function).

This Study

It is difficult to determine the usefulness of the statistical significance tests reviewed in the previous section for equating function selection based on how they were originally proposed. The proposal for using the Kolmogorov-Smirnov and regression tests did not give clear guidelines for how to use the tests in specific equating situations (Budescu, 1987; Jaeger, 1981). The likelihood ratio tests and the equated score difference tests were demonstrated on a few very large datasets (Dorans & Lawrence, 1990; Hanson, 1996;; Moses et al., 2007; von Davier et al., 2004). This study evaluates the previously proposed statistical significance tests in a series of simulations where population equating functions and sample sizes were varied.

Method

Study Design

Simulations were used to study the long-run accuracies of the nine statistical significance tests of interest across two population equating functions and four sample-size conditions. First, population test score distributions and equating functions were defined from test score data obtained in two large-volume equivalent-groups administrations of two tests, *X* and *Y*. Then 200 *X* and *Y* datasets were randomly drawn from each of the two population distributions for four different sample sizes, and the nine statistical significance tests were conducted in each of these 200 X 2 X 4 = 1,600 datasets. Selection rates were computed as the rates at which the significance tests selected the more complex of the two equating functions being compared. Because the population equating functions were known, the selection rates were accuracy rates, indicating tests' power when the tests correctly selected a complex equating function that was the population equating function and indicating tests' Type I error when the tests incorrectly selected an equating function that was more complex than the population equating function.

Nine statistical significance tests. The nine statistical significance tests were those reviewed in the introduction of this paper. In this section they are listed in terms of the two equating functions being compared rather than based on the introduction's ordering of tests' methodological basis.

For comparing the equipercentile and linear equating functions, the three significance tests are:

1. the likelihood ratio test comparing the full loglinear model (4) to the MSD loglinear model (3),

2. the regression test to assess the equipercentile function's curvilinearity (15), and
3. the assessment of equated score differences between the equipercentile and linear equating functions (12) using the Bonferroni procedure.

For comparing the equipercentile and identity equating functions, the three significance tests are:

1. the likelihood ratio test comparing the full loglinear model (4) to the reduced loglinear model (2),
2. the KS test in (8), and
3. the assessment of equated score differences between the equipercentile and identity equating functions (11) using the Bonferroni procedure.

For comparing the linear and identity equating functions, the three significance tests are:

1. the likelihood ratio test comparing the MSD loglinear model (3) and the reduced loglinear model (2),
2. the $\text{KS}_{\text{Linear}}$ test in (9), and
3. the assessment of equated score differences between the linear and identity equating functions (10) using the Bonferroni procedure.

For the study, each significance test was implemented with a defined Type I error rate of 0.05, the likelihood ratio tests were performed by selecting *D* from a range of 3 through 10, and the tests of equated score differences were based on linear equating functions, smoothed equipercentile equating functions where the smoothing was the full loglinear model (4) based on the *D* selected from the likelihood ratio tests, and linear and equipercentile equating functions' conditional standard errors [$SEE_{Y,L}(x_j)$ & $SEE_{Y,E}(x_j)$] and standard errors of equated score differences [$SEED_{Y,E-L}(x_j)$] estimated based on the delta method (Moses & Holland, 2008; von Davier et al., 2004).

Population distributions and equating functions. The populations used in this study were based on two equivalent groups large-volume (≈ 200,000 examinees taking *X* and *Y*) exam administrations where the identity and equipercentile *X*-to-*Y* equating functions were chosen as satisfactory equating functions. From these data, full (4) and reduced (2) loglinear models were fit with $D = 8$, and the X and Y relative frequency distributions from these models were used as the population distributions for the study. Characteristics of X and Y distributions from the full loglinear model (which produced the population equipercentile equating function) and from the reduced loglinear model (which produced the population identity equating function) are shown in Table 1. The X and Y relative frequency distributions are plotted in Figures 1 and 2.

Table 1

Descriptive Statistics for the Population Distributions Used in the Study

Population equating function	Form	Minimum possible score	Maximum possible score	Mean	Std. dev.	Skew	Kurtosis
Identity	X	0	78	43.96	13.35	-0.12	-0.57
	Y	0	78	43.96	13.35	-0.12	-0.57
Equipercentile	X	0	53	32.72	10.08	-0.25	-0.64
	Y	0	53	32.49	10.01	-0.17	-0.72

The X-to-Y equipercentile, linear, and identity equating functions were computed from the population distributions. Difference plots between the three possible equating functions are plotted for the identity population equating condition in Figure 3, and for the equipercentile population equating condition in Figures 4–6. Figure 3 shows that for the identity population equating condition, the differences between the equipercentile, linear and identity equating functions are essentially zero. Figures 4–6 show that for the equipercentile population equating condition, there are differences between the three equating functions across X's score range.

Sample sizes. Random samples of equal sizes of X and Y data were drawn from the X and Y population distributions. The X and Y sample sizes considered were 200, 1,000, 5,000, and 10,000, meaning that, for the simulations, 200 X and Y datasets of size 200, 1,000, 5,000, and 10,000 were drawn from each of the two population distributions.

Accuracy assessments for the statistical significance tests. Accuracy for each significance test was assessed as the rate at which the test selected the more complex of the two equating functions being compared in the 200 datasets. These rates are referred to as *empirical Type I error rates* when the population equating function is the simple identity function and in the sample datasets the significance tests incorrectly select the linear or equipercentile equating function rather than the identity equating function.[2] For Type I error assessments, the statistical hypothesis tests are characterized as *robust* when they do not incorrectly select the more complex

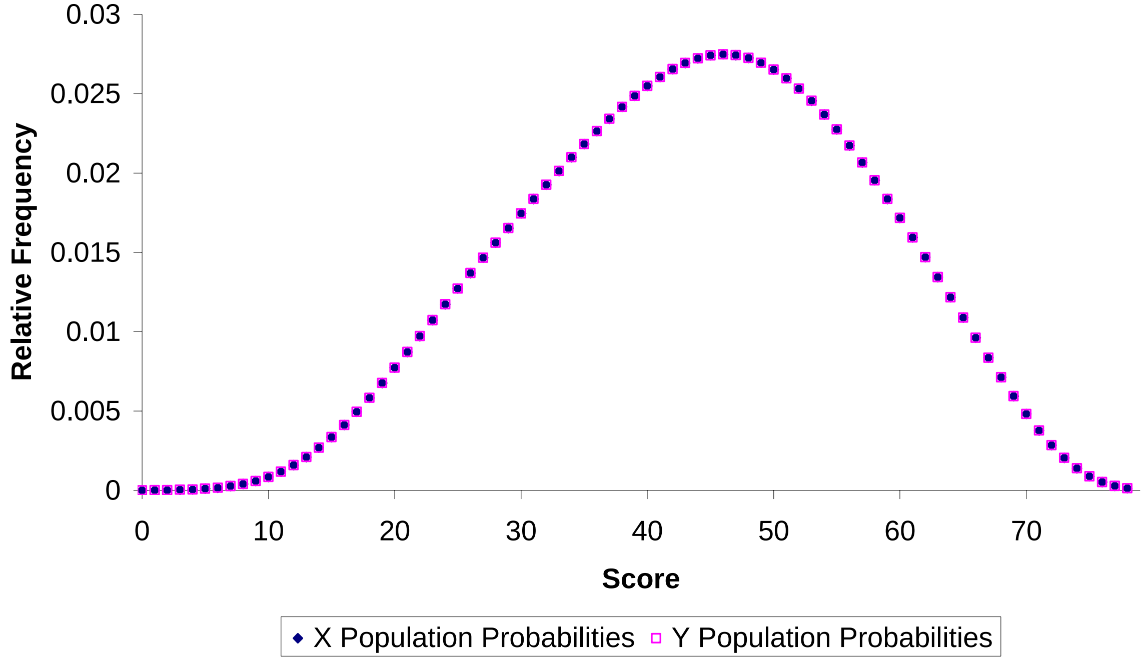

Figure 1. **Population distributions for the identity equating function condition.**

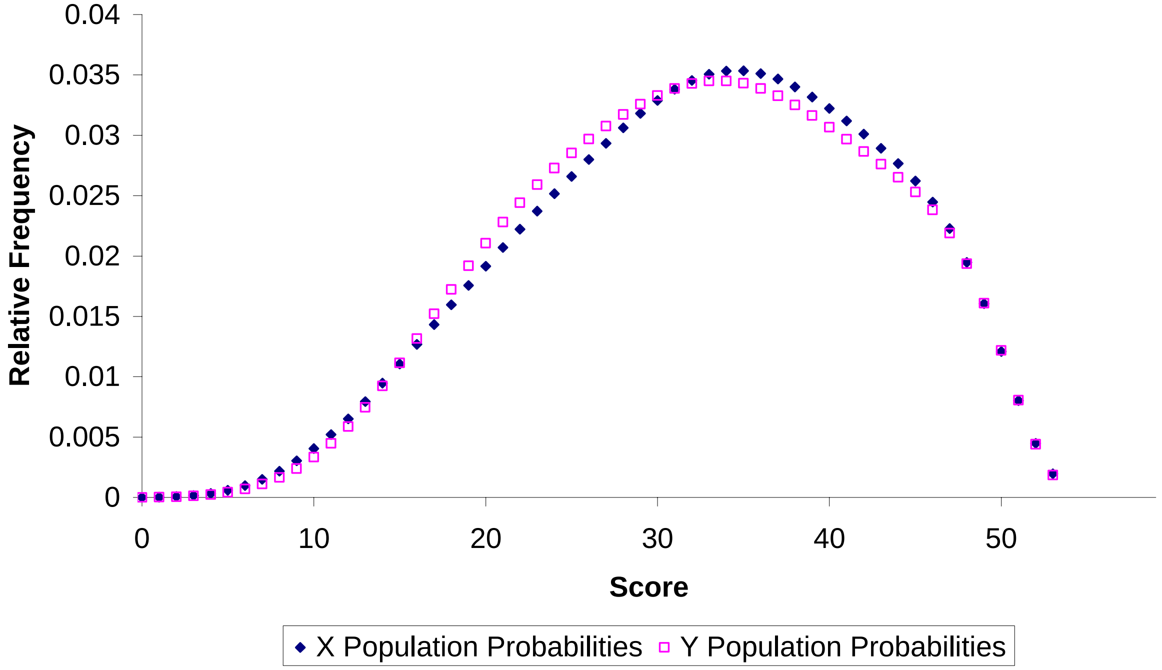

Figure 2. **Population distributions for the equipercentile equating function condition.**

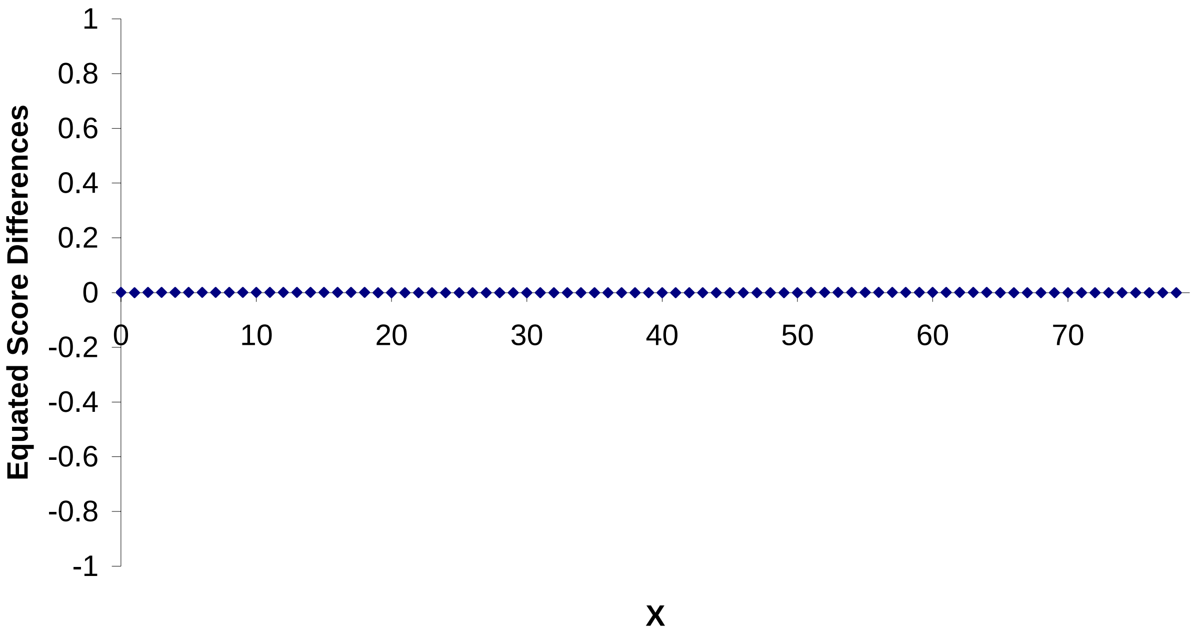

Figure 3. **Equipercentile-linear, equipercentile-identity, and linear-identity equated score differences for the identity population equating function condition.**

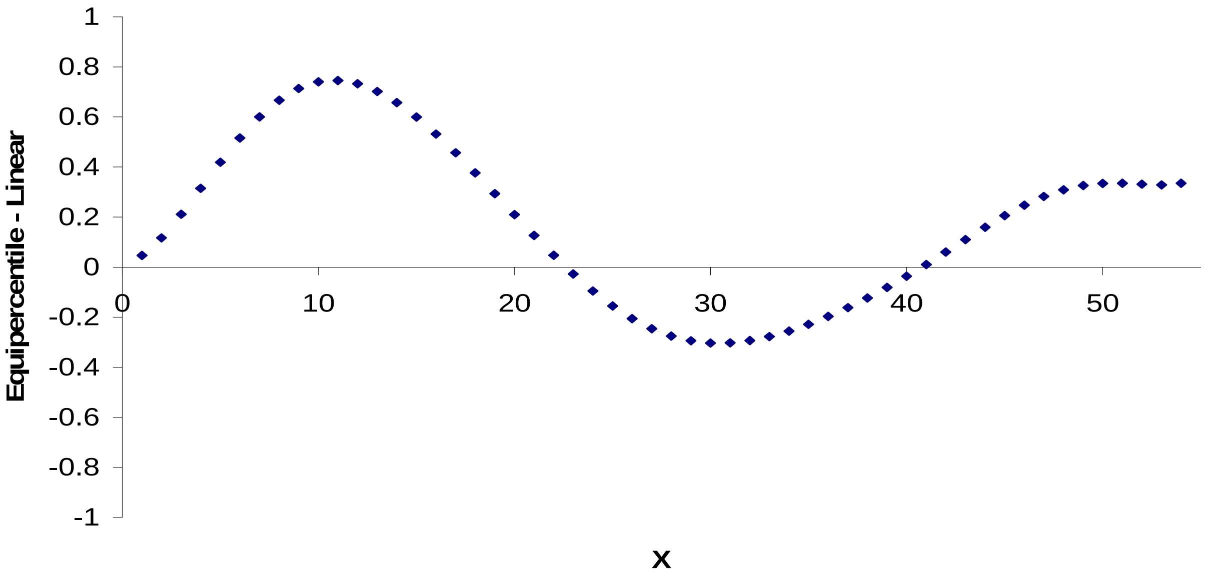

Figure 4. **Equipercentile-linear equated score differences for the equipercentile population equating function condition.**

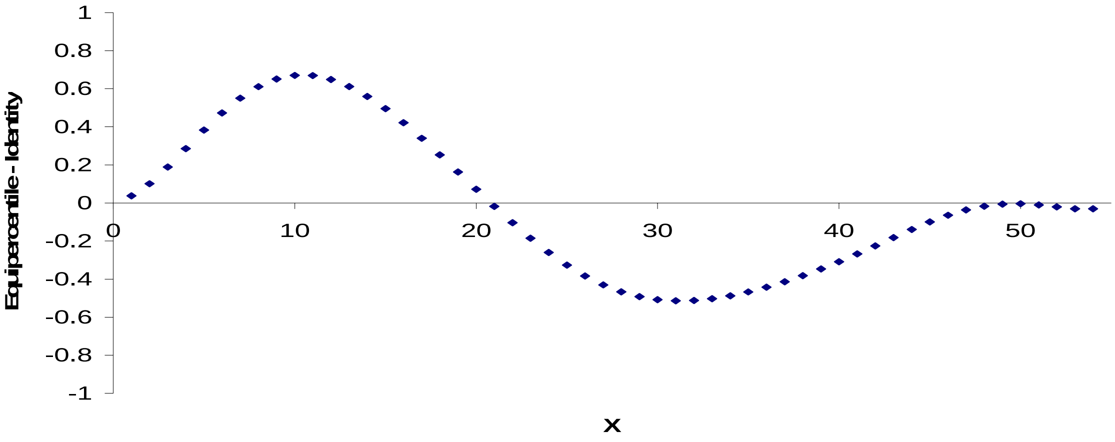

Figure 5. **Equipercentile-identity equated score differences for the equipercentile population equating function condition.**

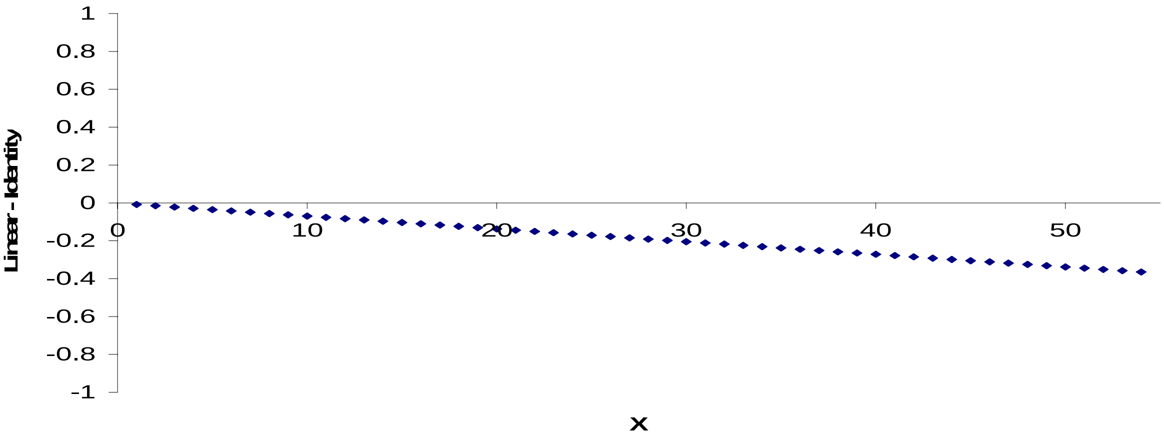

Figure 6. **Linear-identity equated score differences for the equipercentile population equating function condition.**

equating function at a rate that is far above or below the defined Type I error rate (0.05 in this study). Robust and nonrobust empirical Type I error rates for the selection strategies were defined in terms of the +/- 1.96 standard error interval of the defined 0.05 rate, meaning that empirical Type I error rates within the interval of $0.05 \pm 1.96\sqrt{\frac{(0.05)(0.95)}{200\,replications}} = 0.02$ to 0.08 were regarded as robust.

The significance tests' rates of selecting the more complex of the two equating functions being compared are referred to as *power rates* when the population equating function is the complex equipercentile function and in the sample datasets the significance tests correctly select the equipercentile function rather than the simpler linear or identity functions. Significance tests' power rates were computed as the proportion of times they correctly selected this equipercentile equating function in the 200 total datasets. Significance tests' power rates were directly compared to each other to determine the most powerful test.

Some additional Type I error and power assessments were done for the tests of equated score differences (Equations 10, 11, and 12). This study's use of the Bonferroni procedure to produce overall significance tests is somewhat beyond the original development of these tests as score-level significance tests. To further assess the selection rates of the equated score difference tests, plots of their selection rates at each of the individual score levels are provided, some of which indicate score-level Type I error rates and others which indicate score-level power rates.

Results

Type I Error Results

Table 2 presents the empirical Type I error rates for the nine statistical significance tests across the four sample sizes when the population equating function was the identity function. The Type I error rates for the three applications of the likelihood ratio test for comparing loglinear models are robust except for the slightly elevated Type I error (0.09) for comparing full (equipercentile equating function) and MSD (linear equating function) loglinear models with sample sizes of 10,000. The regression test is not robust, in that it incorrectly selected the equipercentile equating function over the linear equating function in 100% of the sample datasets. The tests of equated score differences had Type I error rates that were robust for 6 of the 12 considered equating function comparisons and sample sizes, but were low relative to the defined 0.05 Type I error rate for comparisons of equipercentile and linear equating functions with samples of 200 and 1,000, and for comparisons of linear and identity equating functions with samples of 200, 1,000 and 5,000. The KS test had robust empirical Type I error rates. The KS_{Linear} test had Type I error rates of zero across the considered sample sizes.

Power Results

Table 3 presents the power rates for equating function selection when the population equating function was the equipercentile function. The most powerful strategy was the regression test, which selected the equipercentile equating function over the linear function in 100% of the sample datasets. For the other considered significance tests, the applications of the likelihood ratio test were usually more powerful than the tests for equated score differences. For the comparison of equipercentile and identity equating functions, the KS test was more powerful than the test for equated score differences but was less powerful than the likelihood ratio test. For the comparison of linear and identity equating functions, the $\mathrm{KS}_{\mathrm{Linear}}$ test was less powerful than the test for equated score differences and the likelihood ratio test.

Score-Level Type I Error Rates of Test of Equated Score Differences

Figures 7–9 plot the score-level empirical Type I error rates of the tests for equated score differences based on the 200 datasets for sample sizes of 200 and 10,000. The non-plotted Type I error rates based on sample sizes of 1,000 and 5,000 were very similar to the plotted power rates based on sample sizes of 200 and 10,000. These three plots consider equating function selections when the population equating function is identity, so that the population equating function differences between the equipercentile, linear, and identity functions are (nearly) zero across the *X* scores. Selection rates for the more complex equating function (i.e., the incorrect equating function) in the 200 datasets are plotted for each score in the *X* score range. The dashed line denotes the defined 0.05 Type I error rate.

The results of Figures 7–9 show that the empirical Type I error rates based on the tests of equated score differences across the *X* score range and three equating function comparisons are generally close to the defined 0.05 Type I error rate. The differences in the empirical Type I error rates based on equating samples of 200 or 10,000 are usually not large, though for the comparisons of the equipercentile and linear functions (Figure 7) the rates based on samples of 10,000 are often larger than the defined Type I error rates and those based on samples of 200. Some of the empirical Type I error rates at the lowest and highest *X* scores are higher than the defined 0.05 Type I error rate, mostly for the comparison of the equipercentile and identity equating functions (Figure 8). The score-level Type I error rates exhibit the least fluctuation from

Table 2

Equating Function Selection Rates: Identity Population Equating Function

	Rate at which the equipercentile equating function is selected over the linear equating function (Type I error rates[2])			Rate at which the equipercentile equating function is selected over the identity equating function (Type I error rates)			Rate at which linear equating function is selected over the identity equating function (Type I error rates)		
N_X = N_Y =	Likelihood ratio test	Regression test	Equated score differences	Likelihood ratio test	KS	Equated score differences	Likelihood ratio test	$\mathrm{KS}_{\mathrm{Linear}}$	Equated score differences
200	0.040	0.995[a]	0.010[a]	0.055	0.030	0.085[a]	0.040	0.000[a]	0.005[a]
1,000	0.040	1.000[a]	0.005[a]	0.035	0.025	0.035	0.040	0.000[a]	0.000[a]
5,000	0.065	1.000[a]	0.045	0.050	0.040	0.040	0.055	0.000[a]	0.000[a]
10,000	0.090[a]	1.000[a]	0.045	0.080	0.035	0.060	0.065	0.000[a]	0.065

Note. KS = Kolmogorov-Smirnov distances.

[a] Nonrobust Type I error rates are those that exceed the band of +/- 1.96 standard errors from the nominal 0.05 rate.

Table 3

Equating Function Selection Rates: Equipercentile Population Equating Function

	Rate at which the equipercentile equating function is selected over the linear equating function (Power rates)			Rate at which the equipercentile equating function is selected over the identity equating function (Power rates)			Rate at which linear equating function is selected over the identity equating function (Power rates[2])		
N_X = N_Y =	Likelihood ratio test	Regression test	Equated score differences	Likelihood ratio test	KS	Equated score differences	Likelihood ratio test	$\mathrm{KS}_{\mathrm{Linear}}$	Equated score differences
200	0.075	1.000	0.020	0.055	0.020	0.075	0.070	0.000	0.010
1,000	0.160	1.000	0.060	0.150	0.070	0.060	0.055	0.000	0.000
5,000	0.525	1.000	0.490	0.590	0.335	0.235	0.270	0.020	0.060
10,000	0.730	1.000	0.740	0.810	0.580	0.465	0.385	0.100	0.125

Note. KS = Kolmogorov-Smirnov distances.

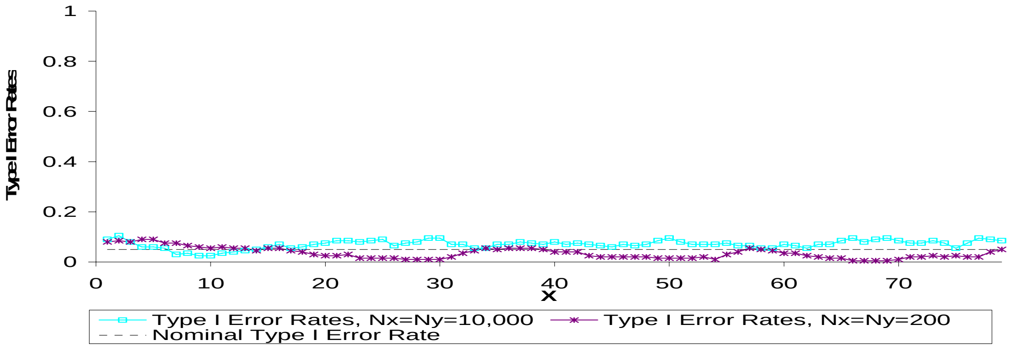

Figure 7. **Type I error rates for score-level tests of equipercentile-linear equated score differences (identity population equating function condition).**

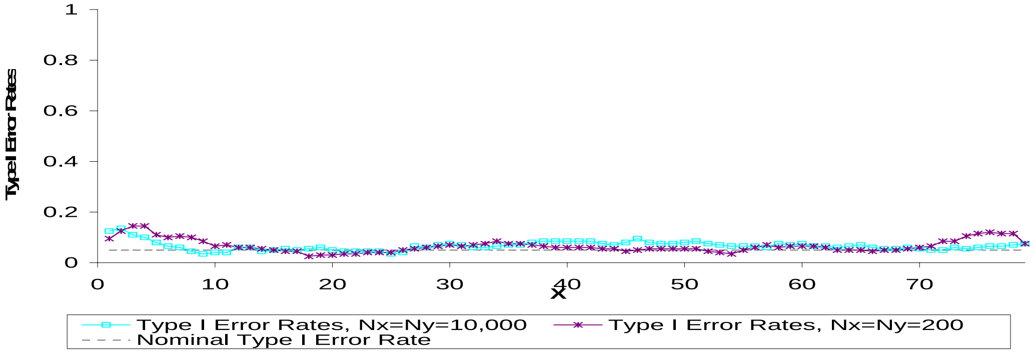

Figure 8. **Type I error rates for score-level tests of equipercentile-identity equated score differences (identity population equating function condition).**

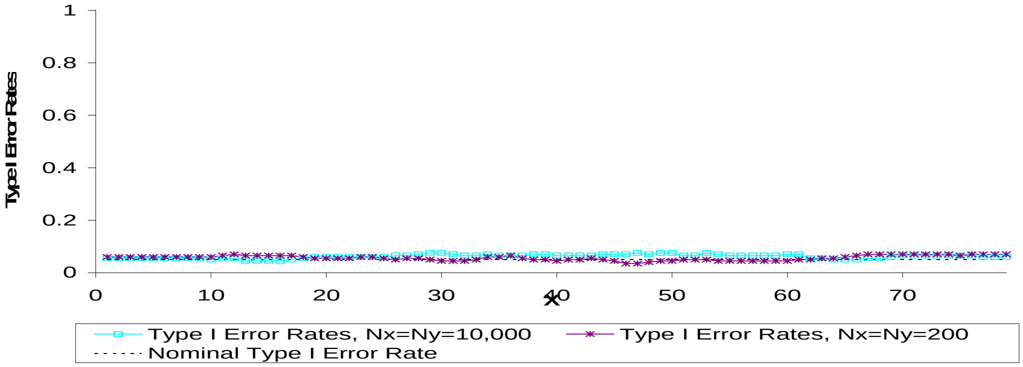

Figure 9. **Type I error rates for score-level tests of linear-identity equated score differences (identity population equating function condition).**

the defined 0.05 Type I error rate for the comparison of the linear and identity equating functions (Figure 9).

Score-Level Power Rates of Test of Equated Score Differences

The power rates for the score-level tests of equated score differences for the population equipercentile equating function are plotted in Figures 10–12 based on testing equated score differences in the 200 datasets for sample sizes of 200 and of 10,000. The non-plotted power rates based on sample sizes of 1,000 and 5,000 were in-between the plotted power rates based on sample sizes of 200 and 10,000. The power rates of the score-level tests of equated score differences in Figures 10–12 are influenced by three factors: overall sample size (either 200 or 10,000), the score ranges likely to be sparse and non-sparse in the datasets sampled from the population distributions (Figure 2), and the magnitude of the equated score differences in the population (Figures 4–6). In terms of overall sample size, the power rates based on sample sizes of 10,000 are much greater than the power rates based on sample sizes of 200. In terms of score-level sample sizes in the datasets sampled from the population distributions, the *X* scores where the power rates are smallest are those where there are likely to be less data (i.e., the *X* scores of 0–7 are likely to have the least data, Figure 2).

In terms of the magnitude of population equated score differences, power rates are high for the *X* scores where the magnitudes of the population equated score differences are large in absolute value:

1. The highest power rates for comparing the equipercentile and linear equating functions plotted in Figure 10 correspond to large population equated score differences shown around *X* scores of 10, 30, and 50 (Figure 4).

2. The highest power rates for comparing the equipercentile and identity equating functions plotted in Figure 11 correspond to large population equated score differences shown around *X* scores of 10 and 30 (Figure 5).

3. The highest power rates for comparing the linear and identity equating functions plotted in Figure 12 correspond to large population equated score differences around *X* scores of 50 (Figure 6). The low power rates in Figure 12 correspond to *X* scores less than 30 where the population equated score differences are near zero (Figure 6).

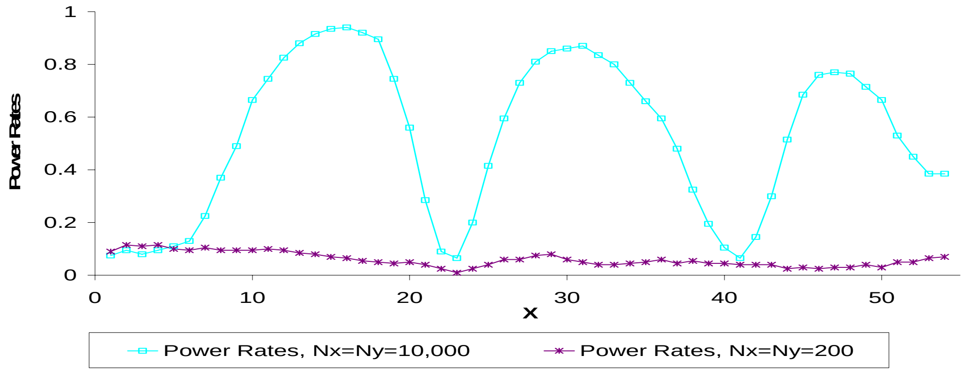

Figure 10. **Power rates for score-level tests of equipercentile-linear equated score differences (equipercentile population equating function condition).**

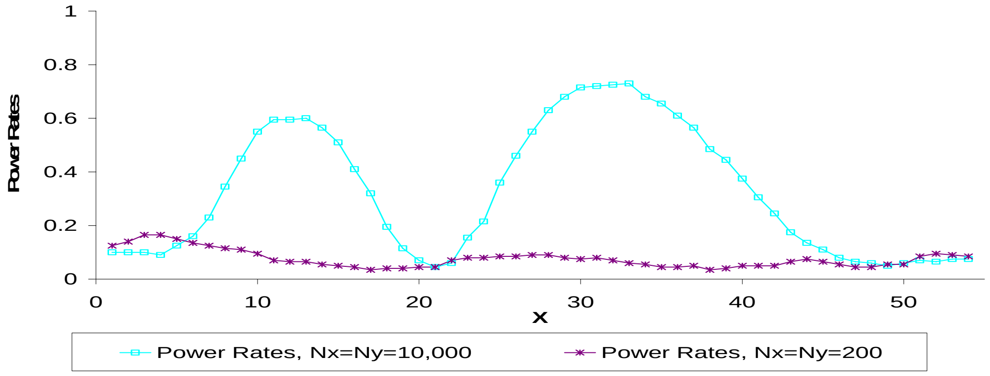

Figure 11. **Power rates for score-level tests of equipercentile-identity equated score differences (equipercentile population equating function condition).**

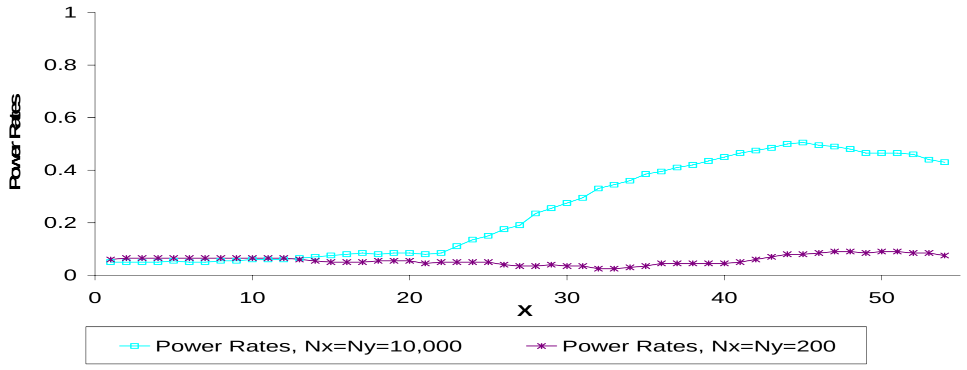

Figure 12. **Power rates for score-level tests of linear-identity equated score differences (equipercentile population equating function condition).**

The results in Figures 10–12 show that at the *X* scores where samples are large (10,000) and the population equated score differences are also large (Figures 4–6), the power for testing equated score differences can be considerably larger than overall significance tests that utilize the Bonferroni procedure (Table 3).

Discussion

Summary

The purpose of this study was to compare several statistical significance tests proposed for selecting equating functions in a simulation that varied the population equating function and the sample sizes. Results showed that the likelihood ratio tests for comparing loglinear models were the most accurate across all of the conditions considered in the simulations. The significance tests of equated score differences were fairly accurate as score-level significance tests when based on large sample sizes, but their use as overall tests and as score-level tests based on small sample sizes suffered from relatively low statistical power. The regression test was not discriminating for assessing curvilinearity in the equipercentile equating functions. Therefore it is not recommended for practice. The Kolmogorov-Smirnov distance test for comparing equipercentile and identity equating functions was fairly accurate, though not as powerful as the likelihood ratio test. This study's application of constraints to the Kolmogorov-Smirnov test for comparing linear and identity equating functions was not accurate enough to be recommended for practice.

Issues With Each Significance Test

The likelihood ratio tests of loglinear models turned out to be the most accurate overall significance tests considered in this study. This was possibly due to the chi-square tests making more efficient use out of all the test score data, in that the likelihood ratio chi-squares (5) took the number of score-level observations into account in ways that the tests for equated score differences (Equations 10, 11, and 12) and the Kolmogorov-Smirnov distances (Equations 8 and 9) did not. Possible drawbacks to the likelihood ratio tests are that the multiple significance tests needed for selecting the overall *D* can be cumbersome, the magnitude of equated score differences cannot be directly observed from the likelihood ratio statistics, and the comparison of equating functions from data-collection designs that utilize bivariate test score data is not feasible. Useful aspects of the likelihood ratio tests are that the equating function comparisons can be understood in terms of differences in test distributions' moments so that additional equating functions can be considered (e.g., Kolen & Brennan's mean equating, 2004), and that the modeled frequencies from the loglinear models can be directly used and relatable to smoothed equating functions.

One difficulty with tests of equated score differences is that they do not directly apply to overall assessments of equating functions. This study used the Bonferroni procedure (Lomax, 2001) to create overall tests out of the equated score difference tests and produced significance tests that were not as powerful as other significance tests. The application of the Bonferroni procedure was least powerful for the comparison of the linear and identity equating functions, an understandable result because the score-level equated scores and standard errors of the linear function do not create the independent significance tests for which the Bonferroni procedure is most justifiable. Other ways to produce overall significance tests out of the tests for equated score differences include alternative familywise Type I error rate procedures or extensions of the delta method (von Davier et al,, 2004) that might produce a summed test statistic out of the score-level tests. Some of these alternatives were considered in preliminary versions of this study but were found to be less accurate than the reported results based on the Bonferroni procedure. The most important finding of this study is that the equated score difference tests were reasonably accurate as large-sample, score-level significance tests but probably do not create the best significance tests for selecting equating functions at an overall level.

The regression test was not very accurate for assessing the extent of curvilinearity in the equipercentile equating function, and is therefore not recommended for practice. The specific problem with the regression appears to be the denominator of the *F* statistic (Equation 15), which is usually based on a linear regression function that closely fits the equipercentile function so that the test statistic gets divided by an extremely small number. The finding that the regression test is not very discriminating was hinted at in Jaeger (1981), but not shown in a systematic simulation study.

The Kolmogorov-Smirnov distance test had accuracy levels for comparing equipercentile and identity equating functions that suggest it is useful for practice for the assessment of raw equipercentile and identity equating function differences. Previous studies of this test emphasized its use as a measure rather than its use as a statistical significance test (Budescu, 1987; Jaeger, 1981), but in this study its Type I error rate was accurate and its power was better than that of the overall test of equated score differences. This study's extension of the Kolmogorov-Smirnov test to produce a comparison of linear and identity equating functions was unsuccessful (as were other attempts to produce a KS_{Linear} test by preserving means and variances in the continuized densities of kernel equating). It should be noted that there are additional variations of this test that could be considered, such as test statistics formed from item-level difficulty or discrimination statistics (Jaeger, 1981), but these comparisons are less relatable to observed score equating comparisons that were the focus of this study.

Suggestions for Practice

The recommendation for applying this study's results to equivalent-groups equating is that the likelihood ratio tests of loglinear models and the equated score difference tests be used together to assess equating function differences at overall levels and also at score levels. This recommendation emphasizes the most accurate features of the significance tests and also encourages a consideration of the magnitude of equated score differences with respect to score-reporting practices. Testing programs may use equating methods for maintaining comparability in a wide range of reported scores or for maintaining comparability in a small number of cut scores, and both these uses of equating could be supported by accurate overall significance tests and score-level tests.

Situations can arise in equating practice where test volumes are so small that the statistical significance tests considered in this study would have unacceptably low power rates for equating function selection. If there was a belief that the test forms being equated did in fact differ in difficulty, one possible response to low sample sizes would be to use the statistical significance tests by defining a large Type I error rate (e.g., 0.10, 0.20, or larger) rather than the 0.05 Type I error rate that is typically used (Dorans & Lawrence, 1990; Hanson, 1996; Moses et al., 2007; von Davier et al., 2004). While manipulating Type I error rates could increase statistical power, it would not completely resolve undesirable consequences of equating tests with insufficient data such as the limited ability to evaluate practically large equated score differences and inaccurate estimates of standard errors of equating. In terms of sample sizes, the limits of statistical significance tests' accuracies correspond to the limits of equating function accuracy, and approaches to addressing these limits tend to rely on assumptions that are not well-supported by observed data.

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.

Budescu, D. (1987). Selecting an equating method: Linear or equipercentile? *Journal of Educational Statistics*, 12, 33–43.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245–254.

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

Hanson, B. A. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education*, *9*(4), 305–321.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). 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.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement*, *18*, 23–38.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lomax, R. G. (2001). Statistical concepts: *A second course for education and the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Moses, T., & Holland, P. W. (2008). *Notes on a general framework for observed score equating* (ETS Research Rep. No. RR-08-59). Princeton, NJ: ETS.

Moses, T., Yang, W., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement*, *44*(2), 157–178.

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, *19*, 279–281.

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

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.

Notes

[1] One difficulty with evaluating the probabilities of the *J* test statistics from (10), (11), and (12) is that there is ambiguity in determining the distribution on which to reference the test statistics. The use of sample-estimated standard errors suggests that the *t*-distribution is appropriate. However, the *t*-distribution requires the specification of degrees of freedom, and there is not a direct way to determine the statistics' degrees of freedom at individual scores when the parameters involved in the equating have been estimated from a loglinear model of the distribution of the *J* frequencies. The use of the *z*-distribution to compute the test statistics' probabilities avoids the question of degrees of freedom but is most appropriate when the standard error is known rather than estimated in the samples. With the large sample sizes usually hoped for in equating, the difference between using the *t*- and *z*-distributions is small, so that the ambiguous choice of the referencing distribution for the test statistics has less impact on the final results.

[2] Some of the comparisons of equating functions do not precisely correspond to Type I error rates or to power rates because neither of the two equating functions being compared are the population equating function. These comparisons are of the linear and identity equating functions when the population equating function is the equipercentile function, and the equipercentile and linear functions when the population equating function is the identity function.