Research Report

Development of Approximations to Population Invariance Indices

Jinghua Liu

Xiaowen Zhu

July 2008

ETS RR-08-36

Development of Approximations to Population Invariance Indices

Jinghua Liu
ETS, Princeton, NJ

Xiaowen Zhu
University of Pittsburg, PA

July 2008

Abstract

The purpose of this paper is to explore methods to approximate population invariance without conducting multiple linkings for subpopulations. Under the single group or equivalent groups design, no linking needs to be performed for the parallel-linear system linking functions. The unequated raw score information can be used as an approximation. For other linking functions that are nonparallel-linear, linking only needs to be conducted for the total population. The difference of the standardized mean differences between each subpopulation and the total population across the old form and the new form can be used as an approximation of population invariance. Under the nonequivalent groups with anchor test design, conducting separate subpopulation linking and comparing them to the total population linking may still be the best way to estimate population invariance.

Key words: Population invariance indices, single group linking, equivalent groups linking, NEAT linking, subpopulation linking

Table of Contents

Page

1. Dorans-Holland Measures of the Population Sensitivity of Score Linking Functions 2

2. Parallel-Linear System of Linking Functions in the SG or EG Design—No Need to Conduct Any Linkings 5

3. Equipercentile Linking in SG or EG Design—Conducting Linking Based on Total Population Only 6

3.1 Results Based on Total Population Linking and Subpopulation Linkings—Full Equatability Analysis 7

3.2 Results Based on Total Population Linking Only—Difference in Standardized Mean Differences Across Verbal and Critical Reading 8

4. Sensitivity Indices in the NEAT Design 10

4.1 Results Based on Full Equatability Analysis in an NEAT Design 11

4.2 Results Based on Approximation: The Difference of the Standardized Mean Differences Between the Total Test and the Anchor Across the Old Form and the New Form 13

5. Discussion 15

References 18

List of Tables

Page

Table 1. Summary Statistics of Full Equatability in an Equivalent Groups (EG) Design 8

Table 2. Difference of the Standardized Mean Differences across Verbal and Critical Reading for Gender Groups 9

Table 3. Sample Sizes for Equating New Form *X* to Old Form *Y* in a Nonequivalent Groups Anchor Test (NEAT) Design 11

Table 4. Summary Statistics of Full Equatability in a Nonequivalent Groups Anchor Test (NEAT)Design 12

Table 5. Raw Score Summary Statistics of Group Performance in a Nonequivalent Groups Anchor Test (NEAT) Design 13

Table 6. Difference of the Standardized Mean Differences Across the Total Test and the Anchor on the Old Form 14

Table 7. Difference of the Standardized Mean Differences Across the Total Test and the Anchor on the New Form.......... 14

Table 8. Comparison of the Difference of the Standardized Mean Differences Between the Total Test and the Anchor Across the Old and the New Forms 15

The goal of test score equating is to ensure that scores from one test form can be used interchangeably with scores from another form. Equating is the strongest form of test score linking. It requires strong assumptions: the same construct requirement, the equity requirement, the symmetry requirement, the equal (and high) reliability requirement, and the population invariance requirement (for more details, see Dorans & Holland, 2000; Holland & Dorans, 2006; Liu & Walker, 2007). In this paper, we focus on the last requirement; the population invariance requirement.

The assumption of population invariance requires that the score equating function should be invariant across subpopulations from the total population from which the subpopulations are drawn. In other words, the equating function ought to be subpopulation independent. Kolen (2004) reviewed the research on population invariance, and concluded that population invariance holds approximately when alternate test forms are built to the same, or very similar, content and difficulty specifications. Equating should be population invariant, while other types of linking are not expected to be invariant (Holland & Dorans, 2006).

The ETS research report titled *Population Invariance of Score Linking: Theory and Application to Advanced Placement Program Examination* (Dorans, 2003); the spring 2004 special issue of *Journal of Educational Measurement*, titled *Assessing the Population Sensitivity of Equating Functions* (Dorans, 2004a); the 2004 special issue of *Applied Psychological Measurement*, titled *Concordance* (Pommerich & Dorans, 2004); and the 2008 special issue of *Applied Psychological Measurement* titled *Population Invariance* (von Davier & Liu, 2008); all contain collections of articles that study population sensitivity issues from different perspectives, across a variety of testing programs. For example, Yang (2004) examined whether the multiple-choice to composite linking functions in the Advanced Placement Program®, or AP®, exam remain invariant over subgroups defined by region. Yin, Brennan, and Kolen (2004) examined group invariance under concordance conditions between ACT and the Iowa Test of Educational Development (ITED) scores. Liu, Cahn, and Dorans (2006) examined population invariance of linking the revised SAT® to the old SAT, to assess the equatability of the revised SAT scores.

The most commonly used population sensitivity indices were developed by Dorans and Holland (2000), where the total population is assumed to be partitioned into mutually exclusive and exhaustive subpopulations, and linking functions are conducted in the total population and in each subpopulation of interest. However, it can be very time consuming and computer-intensive to conduct a separate linking function for each subpopulation (e.g., number of linkings = subpopulations x linking methods x measures.). For example, Dorans, Liu, Jiang, and Cahn (2006) conducted a score equity assessment (SEA; Dorans, 2004b), which produced estimates of means on the new SAT critical reading and math for gender groups and ethnic groups (White, Black, Asian American, Hispanic, and Other), assuming that the old SAT verbal and math had continued to be used. The number of subgroup linkings and scalings was 196 for critical reading and 210 for math, for a total number of 406. Dorans et al. (2006) assumed that the linking method chosen for the total population was appropriate for each of the subpopulations, which may not necessarily be true. If the researchers also tried multiple linking methods for each subpopulation, the total number of linkings would have been multiplied by 5 or 6.

The goal of this study is to explore ways to approximate meaningful yet easily computed population invariance indices that do not require the creation of multiple subpopulation linking functions. This paper is organized in the following way. Section 1 reviews the Dorans-Holland measures of population sensitivity of score-linking functions. Section 2 discusses the parallel-linear system of linking functions in a single population, where there is no need to perform any actual linkings. Section 3 explores ways to approximate population invariance indices based on the total population linking function for the single group (SG) design or equivalent-groups (EG) design, when the linking functions are nonparallel-linear. Section 4 looks at the difference of the standardized mean differences between the total test and the anchor as an approximation for population invariance in the nonequivalent-groups anchor test (NEAT) design. Finally, section 5 synthesizes these findings.

1. Dorans-Holland Measures of the Population Sensitivity of Score-Linking Functions

Dorans and Holland (2000) developed general population invariance indices of linking functions used for one population, either for a single group or for two groups that are equivalent. von Davier, Holland, and Thayer (2004) extended that work to the nonequivalent groups. Holland and Dorans (2006) synthesized the score-linking sensitivity indices across different linking designs. These methodological developments are pertinent to the present paper.

Linking is usually conducted in the total group to produce a total group linking function and a total group scaling function that place raw scores onto the score reporting scale. To examine population invariance of linking functions, linkings and scalings are produced for each subpopulation of interest as well. The Dorans-Holland indices assume that the total population T

is partitioned into several subpopulations, T_j (j = 1, 2, …). X and Y are the two test forms to be linked. The linking on total population T is denoted by the linking function $e_T(x)$, and $e_{T_j}(x)$ denotes the linking function for subpopulation T_j. Each subpopulation is weighted by its relative frequency, w_j , so that $\sum w_j = 1$. The difference between $e_{T_j}(x) - e_T(x)$ is then computed for each subpopulation.

The first index is the root mean square difference measure, RMSD(x), defined as

$$RMSD(x) = \frac{\sqrt{\sum_j w_j \left[e_{T_j}(x) - e_T(x) \right]^2}}{\sigma_{YT}}. \tag{1}$$

RMSD(x) provides an average across groups at each score level. Another index, root expected mean square difference (REMSD), provides a single number summarizing the values of RMSD(x). REMSD is obtained by averaging RMSD(x):

$$REMSD = \frac{\sqrt{E_T \left\{ \sum_j w_j \left[e_{T_j}(x) - e_T(x) \right]^2 \right\}}}{\sigma_{YT}}, \tag{2}$$

where E_T denotes expectation or average over the score distribution of X in T.

In addition, we can also compute the root expected square difference for each subpopulation, RESD(j) to evaluate how close each subpopulation linking function is to the total population linking function:

$$RESD(j) = \frac{\sqrt{E_{T_j} \left[e_{T_j}(x) - e_T(x) \right]^2}}{\sigma_{YT}}. \tag{3}$$

RESD(j) weights by the relative frequency of new form X in the subpopulation T_j. There is a RESD(j) for each subpopulation.

Note that the Dorans-Holland indices are based on the raw-to-raw linking and the divisor σ_{YT} is used to quantity the differences in standard deviation units. However, we need to keep in mind that a raw-to-raw linking or an equating function is a transformation of raw scores on test

X, to the scale of raw scores on test *Y*. It is usually the first step of a two-step process by which raw scores on test *X* are put onto the reported scale on test *Y*. The second step is to convert the equated raw score of *X* to the reporting scale of *Y*, through a scaling function that maps the raw scores of *Y* to the scale. The first step of raw-to-raw equating function and the second step of scaling function are composed to convert the raw scores of *X* onto the reporting scale of *Y* (Holland & Dorans, 2006). The reported or the scaled scores are the final scores that test users get, and most readers are familiar with and can easily interpret scaled score values (e.g., the College Board 200-to-800 scale).

Researchers have modified the standardized Dorans-Holland indices on the raw score scale and expressed the difference in the scaled score unit (Liu et al., 2006). The population invariance indices in the scaled score unit are then defined as:

$$RMSD(x) = \sqrt{\sum_j w_j \left[s_{T_j}(x) - s_T(x) \right]^2}\ , \tag{4}$$

$$REMSD = \sqrt{E_T \left\{ \sum_j w_j \left[s_{T_j}(x) - s_T(x) \right]^2 \right\}}\ , \tag{5}$$

and

$$RESD(j) = \sqrt{E_{T_j} \left[s_{T_j}(x) - s_T(x) \right]^2}\ , \tag{6}$$

where $s_{T_j}(x)$ is the equating and scaling function or the raw-to-scale conversion based on subpopulation T_j, and $s_T(x)$ is the raw-to-scale conversion based on total population *T*.

In order to evaluate the relative magnitude of the differences between subpopulation linking functions and the total population linking function, Dorans and Feigenbaum (1994) proposed the notion of the score difference that matters (DTM), in the context of SAT linking. On the SAT scale, scores are reported in10-point units. For a given raw score, if the unrounded scaled scores resulting from two separate linkings differ by fewer than 5 points, then the scores should ideally be rounded to the same reported score. Dorans, Holland, Thayer, and Tateneni (2003) adapted the above indices used in SAT practice to other tests and considered the DTM to be half of a score unit for unrounded scores.

As can be seen from the formulae above, all of the calculations are based on total population linking and subpopulation linking functions. However, as we mentioned previously, performing multiple linkings can be very time and computer intensive. So the question remains: Is there a short cut that allows us to assess population invariance?

2. Parallel-Linear System of Linking Functions in the SG or EG Design— No Need to Conduct Any Linkings

Dorans and Holland (2000) examined RMSD(*x*) and REMSD for a special case, which they call the parallel-linear system of linking functions in the SG or EG design. The system of parallel-linear linking functions has the same slope between the subpopulation linking functions and the total population linking function. It only allows intercept differences between subgroup/total group linking functions. RMSD(*x*) and REMSD are equal for the parallel-linear system of linking functions:

$$RMSD\left(x\right) = REMSD = \sqrt{\sum_j w_j \left[\left(\frac{\mu_{YT_j} - \mu_{YT}}{\sigma_{YT}}\right) - \left(\frac{\mu_{XT_j} - \mu_{XT}}{\sigma_{XT}}\right)\right]^2}\ , \qquad (7)$$

where μ_{YT_j}, μ_{YT}, μ_{XT_j}, and μ_{XT} denote the unequated raw score means of *Y* and *X* for subpopulation T_j and total population *T*, and σ_{YT} and σ_{XT} denote the standard deviations of the unequated raw scores of *Y* and *X* for the total population. Therefore, as can be seen from the equation, we can estimate population sensitivity without conducting any linkings.

Dorans and Holland (2000) illustrated the computation of RMSD(*x*) values for the parallel-linear case with several examples. They had two forms, *X* and *Y*, two sets of scores, SAT verbal (SAT-V) and SAT math (SAT-M), on each form (the linkings were based on SAT-V to SAT-V and SAT-M to SAT-M), and three ways of forming subpopulations: gender, language spoken at home, and ethnicity. The results showed very little evidence of population dependence by the parallel linking functions.

Liu and Holland (2008) also used this simplified version of RMSD(*x*) to explore the sensitivity of linking functions on the LSAT subpopulations defined by test-takers' gender, ethnicity, geographic regions, whether they applied to law school, and their law school admission status. Population sensitivity was examined in three different linking situations: linking between completely parallel tests, linking between tests that are not strictly parallel but are of comparable reliability, and linking between completely nonparallel tests. Results showed that linking parallel measures of equal reliability exhibits very little group dependence of linking functions across all the subpopulations studied, whereas the linkage of completely nonparallel tests shows substantial population dependence. Besides the main results of the study, it was shown that this simple version of RMSD(x) is a useful tool to assess population sensitivity, without carrying out the actual linkings.

The beauty of this simplified formula for RMSD is that it is very easy to calculate. In reality, however, we often need to deal with situations that are more complicated than this. In the following section, we try to extend this simple version to a nonparallel-linear linking case: equipercentile linking.

3. Equipercentile Linking in SG or EG Design—
Conducting Linking Based on Total Population Only

The equipercentile linking function is set so that the cumulative distribution function (CDF) of scores on form X converted to form Y scale is equal to the CDF of scores on form Y (Braun & Holland, 1982; Kolen & Brennan, 2004). This nonlinear transformation for total population T can be expressed as:

$$y = Equi_{YT}(x) = G_T^{-1}\left[F_T\left(x\right)\right], \tag{8}$$

where F represents the CDF of X, G is the CDF of Y, and G^{-1} is the inverse of the CDF of Y. The intent is that x and y have the same percentile in total population T. Similarly, for subpopulation $\boldsymbol{T_j}$, the transformation equation is:

$$y = Equi_{YT_j}(x) = G_{T_j}^{-1}\left[F_{T_j}\left(x\right)\right]. \tag{9}$$

When we assume that the two CDFs, $F_T\left(x\right)$ and $G_T\left(y\right)$, have the same shape and only differ in their means and standard deviations, the equipercentile linking function becomes linear linking function, $Lin_{YT}(x)$ (Holland & Dorans, 2006), defined as:

$$Lin_{YT}(x) = \mu_{YT} + \frac{\sigma_{YT}}{\sigma_{XT}}(x - \mu_{XT})\,. \tag{10}$$

Within the equivalent groups design, if the two forms can be equated, it is reasonable to assume that the means in the reported score scale should order various subpopulations in the same or a similar way across the new form and the old form (Holland & Dorans, 2006). In other words, the standardized mean difference for each subgroup should be identical or similar across the new and the old forms,

$$\frac{\mu(ss)_{YT_j} - \mu(ss)_{YT}}{\sigma(ss)_{YT}} = \frac{\mu(ss)_{XT_j} - \mu(ss)_{XT}}{\sigma(ss)_{XT}}\,, \tag{11}$$

where $\mu(ss)$ and $\sigma(ss)$ are the mean and standard deviation of the scaled scores, respectively. As shown in Equation 11, the standardized mean difference is a type of effect size that quantifies the mean differences between two groups in standard deviation units. We just need to perform the linking based on the total population only, and then apply this total-population conversion to each subpopulation to get the summary statistics for each subpopulation. If the above equation does not hold for a particular group or groups, then this can serve as an indicator that the linking might be population dependent.

In order to evaluate whether the above standardized mean difference can be used as an approximation of population invariance, and to explore the relationship between the standardized mean difference and the traditional RMSD(x) and REMSD indices, we examined empirical data from the spring 2003 new SAT field trial for illustration purpose. We first summarized the results based on subpopulation linking, which we call *full equatability analysis*. Then, we presented the results based on the total population linking, by examining the standardized mean differences between each gender subpopulation and the total population across the old version of the test (verbal section) and the new version of the test (critical reading section).

3.1 Results Based on Total Population Linking and Subpopulation Linkings—Full Equatability Analysis

In the 2003 new. SAT field trial, the booklets containing the new critical reading and the booklets containing the old verbal were spiraled, in an effort to yield equivalent groups. The resulting groups who took the new critical reading and the old verbal were deemed to be equivalent (Liu et al., 2006). The critical reading section was then linked to verbal through the EG design for total population using equipercentile linking, and produced a total-group conversion. Equipercentile linking was performed for males and females as well, to yield a male-only conversion and a female-only conversion. We call this kind of analysis a *full equatability analysis*, since it involves both total population linking and subpopulation linkings.

Table 1 presents the results of the full equatability analysis. For the 3,801 males who took the critical reading section, the results showed that they would have received a lower mean (474.9) if the male-only conversion (SGL in the table) had been used in place of the total group conversion (TGL in the table), which yielded a mean of 477.9. The mean difference was -3.4, with a standardized mean difference of -.03. For the 5,374 females, the full equatability analysis indicated that they would have obtained a higher mean (482.8) with a female-only conversion than with the total group conversion (480.4), with a mean difference of 2.3 and the standardized mean difference of .02. The RESD statistics are 3.7 and 2.7 for males and females, respectively. The REMSD value was around 3. These values are all below the DTM of 5, which suggests that the linkage of critical reading to verbal was essentially invariant across males and females.

Table 1

Summary Statistics of Full Equatability in an Equivalent Groups (EG) Design

Group	*N*	Linking	Mean	SD	Mean diff	Std mean diff	RESD
Total	9,194	TGL	479.4	107.8			
Male	3,801	TGL	477.9	111.0			
		SGL	474.9	110.0	-3.4	-.03	3.7
Female	5,374	TGL	480.4	105.3			
		SGL	482.8	105.8	2.3	.02	2.7

Note. RESD = root expected square difference, SGL = subgroup linking, TGL = total group linking.

3.2 Results Based on Total Population Linking Only—
Difference in Standardized Mean Differences Across Verbal and Critical Reading

The results in this section were based on total population linking only: We conducted total population linking, and then applied the conversion to males and females, to get the means and standard deviations for each group. We computed the difference in the standardized mean differences across verbal and critical reading for each of the following comparisons: male versus total, female versus total, and male versus female. We then compared the results to those based on the full equatability analysis.

Table 2 provides the means and standard deviations on the old verbal and on the new critical reading. The differences in standardized mean differences are presented as well. The data shows that on the old verbal the means were 474.8 and 479.3, for the male group and the total group, respectively. The standardized mean difference between the male group and the total group was -.04. This difference was based on 2,283 males out of 5,344 test-takers who took the verbal in the field trial. On the critical reading section, the standardized mean difference was -.01 for the male minus the total group. This difference involved 3,801 males out of 9,194 test-takers who took the critical reading in the field trial. The difference in these two standardized mean differences was -.03 across the two tests. When compared to the equatability results described above, the two methods yielded identical values, -.03.

Table 2

Difference of the Standardized Mean Differences Across Verbal and Critical Reading for Gender Groups

	Verbal			Critical reading			
	N	Mean	SD	*N*	Mean	SD	
Total	5,344	479.3	107.9	9,194	479.4	107.8	
Male	2,283	474.8	110.4	3,801	477.9	111.0	
Female	3,055	482.8	105.8	5,374	480.4	105.3	
		Raw diff	Std diff		Raw diff	Std diff	Diff in std diff (Verbal - CR)
M – T		-4.5	-0.04		-1.4	-0.01	-0.03
F – T		3.4	0.03		1.1	0.01	0.02
M – F		-8.0	-0.07		-2.5	-0.02	-0.05

Note. M – T = Male group – total group, F – T -= female group – total group, M – F = male group – female group.

For the female group, the standardized mean difference of female minus total was .03 on verbal and .01 on critical reading. The difference in the two standardized mean differences across the two tests was .02. This difference involved 3,055 females who took verbal and 5,374 females who took critical reading. Once again, this difference was identical to the results produced by the full equatability analysis.

The difference of the standardized mean differences between old verbal and new critical reading for the male minus female comparison was around .05 in absolute value. Compared to the REMSD value, which was around 3 in scaled score units and .027 in standard deviation units, the difference of the two standardized mean differences was about twice that of the REMSD value, considering rounding errors. It is reasonable in that when there are two subpopulations involved, the expected difference from the total population (-.03 for males and .02 for females) should be one-half the difference between the two subpopulations (-.05).

The same pattern was also observed for the math results (Liu & Dorans, 2004). Hence, we may consider using means and standard deviations to estimate population invariance in the EG or SG design, without actually doing any subpopulation linkings.

4 Sensitivity Indices in the NEAT Design

In the NEAT design, population *P* takes form *X* and anchor *A*, and a different population *Q* takes form *Y* and the same anchor *A*. When examinees with different abilities take different forms across different administrations in the NEAT design, it is more complicated to find a shortcut for assessing population sensitivity. However, the common items that are used to control examinee ability differences might be a place to start. In this paper, we only focus on chained linking with the NEAT design.

Chained linking transforms scores through the following chained stages: First link *X* to *A* on population *P*; then link *A* to *Y* on population *Q*. These two linking functions are then composed to map *X* to *Y* through *A*. The first two stages are more like two SG linkings. Within each SG linking, it is reasonable to assume that the means should order various subpopulations in a same or similar way across the anchor and the total test. If population invariance holds across *X* and *Y*, it is also reasonable to assume that the means should order various subpopulations in a same or similar way across *X* and *Y*, and the anchor should order subpopulations in a same or similar way across the two populations. Hence, the mean differences between the total test and the anchor across the old and the new tests should be close. Any deviation could be a sign of subpopulation dependence.

We can use the difference between the standardized mean differences of the total test and the anchor as an approximation. As shown in Equation 12, each component is actually an effect size, describing the differences in standard deviation units:

$$\frac{\mu_{XP_j} - \mu_{XP}}{\sigma_{XP}} - \frac{\mu_{AP_j} - \mu_{AP}}{\sigma_{AP}} = \frac{\mu_{YQ_j} - \mu_{YQ}}{\sigma_{YQ}} - \frac{\mu_{AQ_j} - \mu_{AQ}}{\sigma_{AQ}}, \tag{12}$$

where μ_{XP_j} , μ_{XP} , μ_{AP_j} , and μ_{AP} denote the raw score means of *X* and *A* on subpopulation P_j and population *P*, and σ_{XP} and σ_{AP} denote the standard deviations of *X* and *A* on *P*. Similarly, μ_{YQ_j} , μ_{YQ} , μ_{AQ_j} and μ_{AQ} denote the raw score means of test *Y* and anchor *A* on subpopulation Q_j and population *Q*; and σ_{YQ} , and σ_{AQ} denote the standard deviations of *Y* and *A* on *Q*.

Again, we examine our hypothesis by comparing the full equatability analyses results, which were based on total population and subpopulation linkings, to the results based on the approximation using standardized mean differences.

4.1 Results Based on Full Equatability Analysis in a NEAT Design

Form *X* was a new SAT critical reading section, and Form *Y* was an old SAT-V section. Forms *X* and *Y* were administered operationally in different SAT administrations. Form *X* was linked to Form *Y*, through an external anchor for the total population and each of the ethnic subpopulations. Table 3 contains sample sizes for the total group and ethnic subgroups. Note that these were the linking samples used when the test was equated, while the samples contained in Table 4 were obtained after equating, and were used to project summary statistics. The White group had relative large sample sizes, whereas other ethnic groups had much smaller sample sizes. The chosen equating function was the chained equipercentile equating using log-linear presmoothed data, for the total group and for each subgroup.

Table 3

Sample Sizes for Equating New Form X to Old Form Y in a Nonequivalent Groups Anchor Test (NEAT) Design

	New form	Old form
Total	6,351	15,746
White	3,928	9,096
Black	444	1,686
Hispanic	520	1,215
Asian American	696	1,405
Other	763	2,344

Table 4 summarizes the results based on the total group linking (TGL) and the subgroup linking (SGL), including the difference of the means based on the total group linking and the subgroup linking (the mean diff), and the RESD statistics.

Table 4

Summary Statistics of Full Equatability in a Nonequivalent Groups Anchor Test (NEAT) Design

Group	*N*	Linking	Mean	SD	Mean diff	Std mean diff	RESD
Total	271,751	TGL	526.3	110.0			
Asian American	32,385	TGL	542.9	113.9			
		SGL	537.1	116.5	-5.8	-.05	8.7
White	166,043	TGL	539.4	101.1			
		SGL	539.8	99.9	0.4	.00	2.1
Other	31,202	TGL	536.1	120.7			
		SGL	537.9	121.2	1.8	.02	3.6
Hispanic	21,617	TGL	473.7	104.3			
		SGL	478.6	106.2	5.0	.05	6.4
Black	20,504	TGL	434.1	100.4			
		SGL	433.7	96.7	-0.4	-.00	4.8

Note. RESD = root expected square difference, SGL = subgroup linking, TGL = total group linking.

The results indicate that the Asian American group would have received a lower mean (537.1) if the Asian American-only conversion had been used in place of the total group conversion, which produced a mean of 542.9, with a difference of 5.8 points. Similarly, the Black group would also have had a lower mean (433.7), if the Black-only conversion had been used. For the White, Other, and Hispanic groups, the subgroup-only conversions would have produced higher means than the total group conversion, with the mean differences being positive. The White and Black groups had the smallest mean differences, 0.4 in absolute value. For other subgroups, the mean differences range from 1.8 to 5.8 in absolute value. The biggest mean difference was found in the Asian American group (-5.8), followed by the Hispanic group (5.0). The RESD statistics concur with the mean differences as expected, in that the Asian American and Hispanic groups had the biggest RESD values: 8.7 for Asian American and 6.4 for

Hispanic. The differences for the Asian American and Hispanic groups were considered large enough (exceeding the DTM) to exhibit group dependence.

In summary, the White group did not exhibit population sensitivity, whereas the Asian American and Hispanic groups exhibited large differences between the subgroup linking and the total group linking, to a degree that merits investigation.

4.2 Results Based on Approximation: The Difference of the Standardized Mean Differences Between the Total Test and the Anchor Across the Old Form and the New Form

This section examines the difference of the standardized mean differences between the total test and the anchor across the old form and the new form, as an approximation. Table 5 contains the raw score summary statistics of population *P* taking form *X* and anchor *A*, and population *Q* taking form *Y* and anchor *A*, broken down by group membership.

Table 5

Raw Score Summary Statistics of Group Performance in a Nonequivalent Groups Anchor Test (NEAT) Design

Group	Old form		New form	
	Total test	Anchor	Total test	Anchor
Test length	78	19	67	19
Total group - mean	37.41	9.21	34.36	9.73
- SD	18.17	4.98	15.56	4.92
Asian American	37.54	9.62	36.47	10.56
	19.56	5.36	16.13	5.21
Black	22.05	5.36	21.17	5.91
	14.88	4.27	15.08	4.67
Hispanic	29.89	6.91	27.09	7.42
	16.52	4.64	14.97	4.83
White	40.67	10.06	36.08	10.23
	16.42	4.52	14.40	4.57
Other	39.60	9.66	36.20	10.24
	19.94	5.41	16.30	5.15

First, we calculated the standardized mean difference for each pair of subgroup minus total group on the total test and on the anchor on the old form *Y*. Table 6 lists the results. For example, the standardized mean difference between the Asian American group and the total group was .01 on the total test, and .08 on the anchor. The difference was -.07. Relatively speaking, the Asian American group did a little worse on the total test than on the anchor. So did the Black group, also with a difference of -.07. The White group did about the same on the total test and on the anchor. The Other group and the Hispanic group did a little better on the total test than on the anchor.

Table 6

Difference of the Standardized Mean Differences Across the Total Test and the Anchor on the Old Form

Group	Old form		Total - anchor
	Total	Anchor	
Asian American	0.01	0.08	-0.07
White	0.18	0.17	0.01
Other	0.12	0.09	0.03
Hispanic	-0.41	-0.46	0.05
Black	-0.84	-0.77	-0.07

Second, we got the standardized mean difference on the total test and on the anchor for each subpopulation on the new form *X*. The results are summarized in Table 7. Again, the Asian American group and the Black group did a little worse on the total test than on the anchor. But the Hispanic group did just about the same on the total test as on the anchor.

Table 7

Difference of the Standardized Mean Differences across the Total Test and the Anchor on the New Form

Group	New form		Total - anchor
	Total	Anchor	
Asian American	0.14	0.17	-0.03
White	0.11	0.10	0.01
Other	0.12	0.10	0.02
Hispanic	-0.47	-0.47	0.00
Black	-0.85	-0.78	-0.07

Third, we compared the (total minus anchor) difference across the old and the new forms. As can be seen from Table 8, the difference was -.04 for the Asian American group, and .05 for the Hispanic group. We also put the full equatability analysis results in Table 8, for the purpose of comparison. As we can see, the results based on the two methodologies are quite similar.

Table 8

Comparison of the Difference of the Standardized Mean Differences Between the Total Test and the Anchor Across the Old and the New Forms

Group	Diff of (total – anchor)		Std. mean diff of (total – anchor) across the old and the new forms	Std. mean diff of (SGL – TGL) based on full equatability analysis
	Old form	New form		
Asian American	-0.07	-0.03	-0.04	-0.05
White	0.01	0.01	0.00	0.00
Other	0.03	0.02	0.02	0.02
Hispanic	0.05	0.00	0.05	0.05
Black	-0.07	-0.07	0.00	-0.00

Note. SGL = subgroup linking, TGL = total group linking.

However, at present, there is some disagreement about using this method. It is argued that *P* and *Q* are two different populations; hence they are not directly comparable (N. Dorans, personal communication, April 23, 2007). It is argued that this method neglects the possible interactions between the group membership and the test difficulty. Even if the difference of total minus anchor standardized mean differences across the old and the new forms is zero for a particular subgroup, it just means that this particular group finds the anchor test being similar to the total test at the difficulty level, in both the old form and the new form, but it does not reveal the relationship between the group membership and the form difficulty across the new form and the old form.

5. Discussion

The purpose of this paper was to explore methods in identifying population invariance, without conducting multiple linkings for subpopulations. Under the SG or EG design, no linking needs to be performed for the parallel-linear system linking functions. The RMSD(x) is equal to the REMSD value that can be calculated using unequated raw score information. For other linking functions that are nonparallel-linear, linkings only need to be conducted for the total population. The total population conversion can then be applied to different subpopulations, and the difference of the standardized mean differences between each pairing of subpopulation and the total population across the old form and the new form can be used as an approximation of the full equatability population invariance indices. However, we would like to point out that the RMSD(x) statistics quantify weighted differences between subgroup versus total group linking functions at each score level, whereas the approach of standardized differences only take into account the means and standardized deviations of equated scores, ignoring the relative frequencies across score levels. Hence, small standardized mean differences cannot warrant population invariance at score levels.

It is more complicated with the NEAT design when it involves two different populations. The difference of standardized mean differences between the total and the anchor test across the old and the new forms might be useful, but there is debate about using it. The results here were only based on one data set. More evidence needs to be collected. In addition, we basically used chained linear linking function, which may not be appropriate to expand to other linking situations where the relationship is not linear.

This paper does not explore alternative ways to calculate population invariance indices with chained curvilinear linking and post stratification linking. These might be topics for future research. For example, it may be possible that we can break down the chained linking into 2 SG linkings, conduct chained curvilinear linking within each SG, and evaluate population invariance in each SG linking, using the standardized mean difference based on total group linking.

In the case of post stratification equating (PSE), such as Tucker linear equating, we can first perform regression of X on A in total population P and in different subpopulations, to get a regression slope and a regression intercept for each subpopulation. We also need to calculate the conditional variance of X given A, in population P and for each subpopulation. If the slopes, intercepts, and conditional variances are invariant across subgroups, then it is likely that the conditional distribution of X given A is population invariant within population P. A similar set of analyses would need to be done within population Q, to determine whether the conditional distribution of Y given A is population invariant. If population invariance is satisfied in both populations, then population invariance is going to hold in the synthetic population, given the assumptions of PSE. However, in this case, the amount of actual work is not reduced. Instead, it gets increased. If we perform a regression analysis for each subgroup in populations P and Q, it is reasonable that we might want to go ahead and conduct the equating for each subgroup.

Essentially, the approximation methods of using standardized mean differences proposed in this study are pretty much based on the assumptions of linear equating or linear linking. It seems paradoxical, though, to evaluate population invariance of nonlinear linking functions using such linearity-based statistics. Therefore, we suggest using the standardized mean difference only as an approximation of population invariance in the SG or EG design. Under the NEAT design, conducting individual subpopulation linkings and comparing them to the total population linking is probably still the best way to determine population invariance.

References

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Dorans, N. J. (Ed.). (2003). *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS Research Rep. No. RR-03-27). Princeton, NJ: ETS.

Dorans, N. J. (2004a). Assessing the population sensitivity of equating functions (Special issue). *Journal of Educational Measurement, 41*(1).

Dorans, N. J. (2004b). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

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, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. 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*(4), 281–306.

Dorans, N. J., Holland, P.W., Thayer, D. T., & Tateneni, K. (2003). 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 Research Rep. No. RR-03-27, pp. 79–118). Princeton, NJ: ETS.

Dorans, N. J., Liu, J., Cahn, M., & Jiang, Y. (2006). *Score equity assessment of transition from SAT I Verbal to SAT Critical Reading: Gender* (ETS Statistical Rep. No. SR-06-61). Princeton, NJ: ETS.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Prager.

Kolen, M. J. (2004). Population invariance in equating and linking: Concepts and history. *Journal of Educational Measurement, 41*(1), 3–14.

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

Liu, J., Cahn, M., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linking of new SAT to old SAT across gender groups. *Journal of Educational Measurement, 43*(2), 113–129.

Liu, J., & Dorans, N. J. (2004). *Projected changes in ethnic and gender group performance: An approximate assessment for the new SAT* (ETS Statistical Rep. No. 2004-23). Princeton, NJ: ETS.

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York: Springer-Verlag.

Liu, M., & Holland, P.W. (2008). Exploring the population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32*(1), 27–44.

Pommerich, M., & Dorans, N. J. (Eds.). (2004). Concordance [Special issue]. *Applied Psychological Measurement, 28*(4).

von Davier, A. A., & Liu, M. (Eds.). (2006). *Population invariance of testing equating and linking: Theory extension and applications across exams* (ETS Research Rep. No. RR-06-31). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods of observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*(1), 15–32.

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). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*(4), 273–289.