Research Report

Effect of Repeaters on Score Equating in a Large-Scale Licensure Test

Sooyeon Kim

Michael E. Walker

July 2009

ETS RR-09-27

Effect of Repeaters on Score Equating in a Large-Scale Licensure Test

Sooyeon Kim and Michael E. Walker
ETS, Princeton, New Jersey

July 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study investigated the subgroup invariance of equating functions for a licensure test in the context of a nonequivalent groups with anchor test (NEAT) design. Examinees who had taken a new, to-be-equated form of the test were divided into three subgroups according to their previous testing experience: (a) repeaters who previously took the reference form, to which the new form would be equated; (b) repeaters who previously took any form other than the reference form; and (c) first-time test takers for whom the new form was the first exposure to the test. Equating functions obtained with two subgroups, all repeaters, and first-time test takers were similar to those obtained with the total group, supporting score equatability of the two forms. However, when the repeater subgroup was divided into two groups based on the particular form examinees took previously, subgroup equating functions substantially differed from the total-group equating function, indicating subgroup dependency of score equating. The results indicate that repeater membership needs to be more clearly specified to assess the impact of repeaters on score equating. Such clarification may be especially necessary for high-stakes licensure tests because repeaters tend to perform more poorly on such tests than first-time test takers.

Key words: Equating, test repeater, population invariance, score equity assessment, licensure test

Acknowledgments

The authors would like to thank Neil Dorans, Daniel Eignor, Mary Grant, and Wen-Ling Yang for their helpful comments and Kim Fryer for the editorial help.

Introduction

In most if not all testing programs with high-stakes consequences, different forms of a single test are used in different administrations to help prevent unfair advantages for those who may have access to information from previous administrations. Although the different test forms are designed to measure the same content and to have the same statistical characteristics, they often differ slightly from one another. To ensure fairness to examinees taking different forms, the scores on the forms are adjusted to make them equivalent.

Test equating is a statistical method of adjusting for differences in difficulty among forms built to the same specifications. One basic requirement for score equating is that equating functions be subpopulation invariant (Dorans & Holland, 2000); that is, the equating function must operate independently of examinees or subgroups of examinees from whom the data are drawn to develop the conversions. If equating functions are not invariant across subpopulations, the new and reference test forms are not equatable and their interchangeability is questionable.

Theoretically, the population invariance requirement means that the equating function must operate independently of subgroups of examinees from whom the data are drawn to develop the conversions (Angoff, 1971). In reality, however, many experts have concluded that all equatings are somewhat population-dependent. Testing organizations often assume invariance of equating functions across different subgroups. This assumption might not be valid in certain situations because the degree of population dependence tends to be test-specific, depending on the definition of the population, the characteristics of the ability being measured, and the linking method used. Thus, the degree of population invariance for a particular test or for particular subgroups (based, for example, on gender, race, or repeater status) cannot be generalized to other tests or to other subgroups for the same test. Accordingly, all testing programs with high-stakes outcomes need to entail periodic investigations for major subgroups that could differ in ways related to the ability being measured, to see if the results are comparable across subgroups.

This study investigated the population invariance of score equating by comparing equating functions derived using two distinct subgroups from the examinee population for a large-scale licensure test. The two subgroups were repeaters (those who had taken the test before) and first-time test takers (those taking the test for the first time, hereafter called *first-timers*). Comparing equating functions for repeaters and first-timers is particularly important for licensure tests because these groups considerably differ in ability on such tests. In general, an

examinee's need to retake the test indicates that he or she did not previously achieve a score high enough for licensure.

The few studies that have examined the impact of repeating examinees on equating have shown inconsistent results (Andrulis, Starr, & Furst, 1978; Cope, 1985; Puhan, 2009). Andrulis et al. compared an equating function derived from the total sample to the function derived from a subsample of first-timers. They found that the performance of repeaters tended to lower the raw scores that map to the passing scores. This effect became more pronounced as the number of repeaters increased. On the basis of this result, Andrulis et al. recommended that repeaters be excluded from the equating conducted to produce the conversion for reporting scores; this conversion then could be applied to the total sample, including both repeaters and first-timers. Cope obtained somewhat different results: equating functions derived using only first-timers did not differ from those derived using the total sample. On the basis of this result, Cope concentrated on the practice of excluding repeaters from the equating sample and called for further research to assess the impact of eliminating large numbers of repeaters on the resulting equating function. Using data from certification tests, Puhan examined the impact of including repeaters in the equating sample on final equating results. The differences between the equatings derived using all examinees and those using only first-timers were very small, with no practical impact on examinees' pass/fail designations.

Excluding repeaters from the equating sample may be reasonable under some circumstances. However, doing so could create an equating group that does not represent the group that was tested, especially if many examinees repeated the test. Excluding repeaters also could create other problems, such as an increase in the standard error of equating due to reduced sample size (Kolen & Brennan, 2004). On the other hand, inclusion of repeaters can cause problems if the repeaters in the new-form sample performed substantially better on the common (i.e., anchor) items than on the remainder of the test because of previous exposure. In such a case the new-form sample would appear to be more proficient than the reference-form sample; consequently, the new form would appear more difficult than it really was. This incorrect adjustment could be substantial if the proportion of repeaters were large.

Perhaps the most commonly used equating design is the nonequivalent groups with anchor test (NEAT) design, in which each test form is administered to one of two groups that differ in ability. Under the NEAT design, the primary rationale for considering repeaters is that

they may tend to perform better on the common items than on the unique new items. We would expect this argument to hold if repeaters were previously exposed to the reference form to which the new form would be equated. However, we would expect repeaters who took any form other than the reference form to perform similarly on both common and unique items because both types of items would be new to these examinees.[1] If this is the case for repeaters who took any form other than the reference form, they should be divided into two groups, producing three subgroups of examinees: (a) first-timers, who never took the test before; (b) repeaters who took the reference form, hereafter called *reference repeaters*, and (c) repeaters who took any form other than the reference form, hereafter called *other-form repeaters*.

A subgroup invariance study that deals with examinees' repeater status as a major subgroup factor offers increased understanding of the impact of sample selection on test score equating. The present study examined subgroup dependency of the equating function using two examinee categories based on repeater status. The first category included two subgroups: (a) first-timers and (b) undifferentiated repeaters. The second category included three subgroups: (a) first-timers, (b) reference repeaters, and (c) other-form repeaters. We assessed the impact of the repeater effect by comparing the equating functions derived from the different subgroups.

Method

Data

Data sets from two national administrations of a large-scale licensure test were used. The data were collected using a NEAT design. The test consisted of 107 multiple-choice items; 42 items were common across new form *X* and reference form *Y*. Descriptive statistics are summarized in Table 1. The table shows that the total new-form group was as adept as the total reference-form group. The mean anchor scores of the two groups differed by only 0.14 correct answers, an effect size of 0.03. These results indicate a negligible difference between the two populations. Therefore, the difference in the mean scores on the two forms (an effect size of -0.36) seems attributable to the difference in difficulty between the two forms. In both the new-form and reference-form groups, first-timers showed higher ability (as measured by the raw test scores) than repeaters, yielding a mean difference larger than one standard deviation. Figure 1 presents the relative frequency distributions, given as percentages of the total group, of new-form scores in the three subgroups.

Table 1

New- and Reference-Form Examinee Groups: Summary Statistics

Test form	Group	*N*	%	Total		Anchor		*r*
				M	*SD*	*M*	*SD*	
New	Total	6,426	100	73.62	10.51	30.60	4.96	.91
Form *X*	First-timer	4,211	66	77.12	9.55	32.16	4.45	.89
	Repeater	2,215	34	66.95	8.88	27.62	4.51	.86
	Reference	330	5	64.66	8.54	26.81	4.57	.86
	Other-form	1,885	29	67.35	8.88	27.76	4.48	.86
Reference	Total	6,489	100	77.47	10.83	30.46	5.09	.92
Form *Y*	First-timer	4,117	63	81.90	9.19	32.49	4.29	.90
	Repeater	2,372	37	69.79	9.00	26.93	4.40	.86
	Reference	470	7	69.26	8.62	26.20	4.21	.85
	Other-form	1,902	29	69.92	9.09	27.12	4.43	.87

Note. Reference repeaters indicate repeaters who took the exact same reference form at previous administrations; *other-form repeaters* indicate repeaters who took any form other than the reference form at previous administrations; *r* is the correlation between the total and anchor scores.

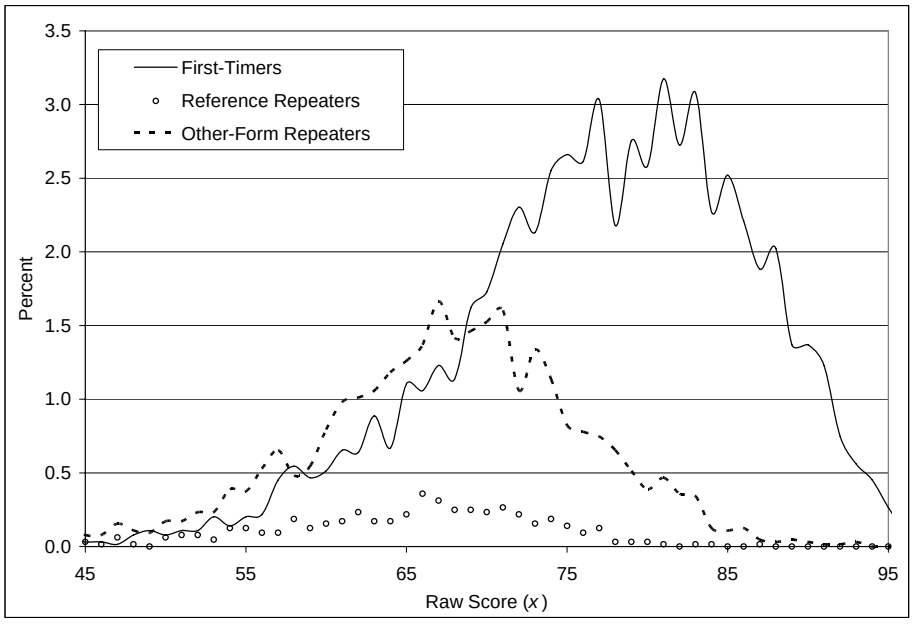

Figure 1. **Relative frequency distribution of new form *x* scores in the three subgroups.**

Procedure

Figure 2 presents the general framework of possible equatings using different subgroups. In the NEAT design, the equating functions derived using either repeaters or first-timers were compared to the equating function derived using all examinees to determine whether the resulting equating function from the total group would produce scores comparable to the subgroup equating functions regardless of the examinees' previous testing experience. The present study included the following three steps.

Step 1: Obtain total-group and subgroup equating functions. The equating relationship between the new (*X*) and reference (*Y*) forms was derived using three groups: (a) total examinees, (b) first-timers, and (c) repeaters. These equating relationships are represented with solid lines in Figure 2. Two additional equating functions were derived using the two subgroups of repeaters: (a) those who took the reference form and (b) those who took any other form. These equating relationships are represented with dashed lines in Figure 2.

Step 2: Compare total-group and subgroup equating functions. The equating function derived using each subgroup configuration was compared to the total-group equating function. The differences were quantified across all subgroups using the root-mean-square difference (RMSD) and root-expected-mean-square difference (REMSD) deviance measures (see Equations 1 and 2). In general, a negligible difference indicates that the equating relationship is not influenced significantly by the subgroups used in deriving that function. In addition, the difference between subgroup equating functions and the total-group equating function was separately quantified for each subgroup to assess more clearly the impact of reporting scores for that subgroup based on the total-group equating transformation (root-expected-square difference, or RESD; see Equation 4). The impact of three different conversions on examinees' pass/fail designations was assessed.

Step 3: Estimate precision of the deviance measures. A total of 2,000 replications were obtained in each equating design using a bootstrap resampling technique (as implemented in SAS PROC SURVEYSELECT) to estimate a 95% confidence band for the RMSD measure conditioned on each raw score. In each replication, examinees were randomly drawn *with replacement* from each reference-form and new-form group until bootstrap samples included the same number of examinees as the actual reference-and new-form groups. Both the reference- and new-form samples then were divided into either two or three mutually exclusive subgroups,

according to examinees' repeater membership. In each replication, three groups (total group, repeater subgroup, first-timer subgroup) were formed in the two subgroup categories. The repeater subgroup was divided into reference repeaters and other-form repeaters in the three-subgroup condition. In each replication, we equated the new-form scores to the reference-form scores for the total group and the two or three subgroups using the chained equipercentile method, and we calculated the RMSD using both equal and proportional (unequal) weights. For the two weighting methods, which method produced the larger RMSD values depended on how the difference in equated scores was paired with the group weight for the subgroups. The 95% confidence interval (CI) for the RMSD measure, which covers RMSDs in the 2.5th–97.5th percentile range, was constructed on the basis of the 2,000 replications.

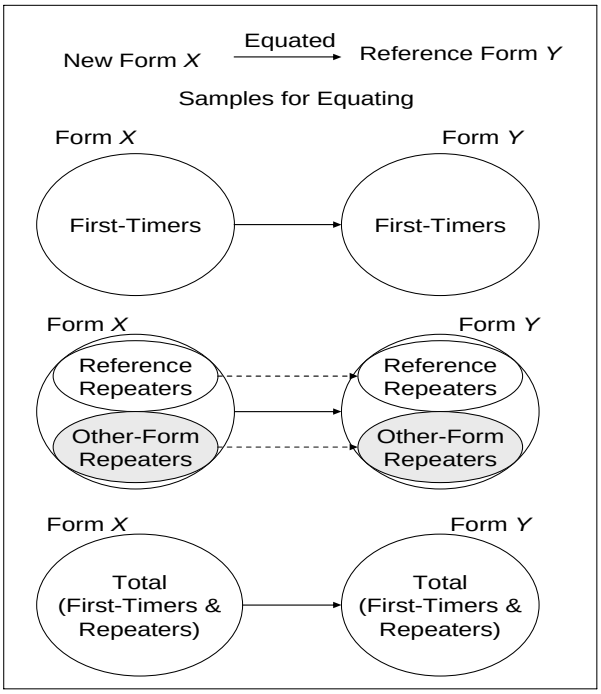

Figure 2. **The general framework for equating new form *X* to reference form *Y*.**

Equating Method

The chained equipercentile equating method was used to derive the equating relationship between the new and reference forms of the test. The data were presmoothed using a log-linear model that preserved the first five univariate moments of each marginal distribution (i.e., of the total score and of the anchor score). No bivariate moments were preserved.[2]

Deviance Measures

Four deviance indices were used as population invariance measures and calculated using the following formulas. von Davier, Holland, and Thayer (2004) defined the RMSD for the anchor test or NEAT data-collection design, and Holland (2003) defined the REMSD. The REMSD index was used to obtain a single value summarizing the values of RMSD(x) over the distribution of x in the total group. The *ew*REMSD (Kolen & Brennan, 2004, p. 443), which gives equal weight to all score points,[3] was also calculated, particularly for the cut-score region (63–72),[4] to elucidate subgroup influence on examinees' pass/fail designations. In addition, the RESD was computed as the weighted average of the squared differences between each subgroup equating function and the total-group equating function at each raw-score level. Thus, each subgroup had a single summary RESD value. We did not standardize any of the measures by dividing by the standard deviation, as is common practice. We were able to use unstandardized measures because we did not conduct any comparisons across different testing programs. The use of raw measures left the deviance indices in the metric of the raw scores and facilitated interpretation of results (see Liu, Feigenbaum, & Dorans, 2005, for another instance of this practice). The various indices are defined as follows.

$$RMSD(x) = \sqrt{\sum_j w_j [e_{yji}(x) - e_{yi}(x)]^2}, \tag{1}$$

$$REMSD = \sqrt{\sum_j w_j \sum_i r_i [e_{yij}(x) - e_{yi}(x)]^2}, \tag{2}$$

$$ewREMSD = \sqrt{\sum_j w_j \sum_i [e_{yij}(x) - e_{yi}(x)]^2}, \tag{3}$$

$$RESD_j = \sqrt{\sum_i r_i [e_{yij}(x) - e_{yi}(x)]^2}, \tag{4}$$

where x represents each raw score point, $e_{yji}(x)$ indicates the equating function in the jth subgroup, $e_{yi}(x)$ represents the equating function in the total group, w_j denotes the proportion of subgroup j in the total group, and r_i indicates the relative proportion of examinees in the total group at each raw-score level.

As shown in Table 1, the subgroup sizes were substantially unbalanced. The proportion of first-timers was over 60%, whereas the proportion of reference repeaters was lower than 10%

in both groups. Thus, the first-timer subgroup heavily influenced both the RMSD and REMSD measures. The RMSD measure itself might be biased against smaller groups (Yang, 2004). As a simple way to adjust for the unbalanced proportions, equal weight, along with the proportional (i.e., unequal) weight derived from the actual relative size of each subgroup, was imposed on each group when calculating RMSD and REMSD (e.g., w_j was always 0.5 in the two-subgroup condition, and w_j was always 0.333 in the three-subgroup condition).

To determine when the REMSD was large enough to warrant concern about the equatability of two forms, the notion of the score *difference that matters* (DTM; Dorans & Feigenbaum, 1994), defined as half of a raw score point in the raw-to-raw score transformations (i.e., 0.5). To evaluate equating difference from a statistical perspective, the 95% CI of the RMSD was generated using a bootstrap resampling technique.

Results

Preliminary Analysis

To clarify the effect on the reference-repeater group of previous exposure to common anchor items, moderated regression analyses were performed that predicted new-item (nonanchor) scores from anchor-item scores, repeater membership (reference vs. other-form), and the interaction of anchor and repeater membership. After this overall analysis, separate regression lines for the two repeater subgroups then were computed to predict new-item scores from the anchor score, by plugging repeater membership information (e.g., 1 for reference repeaters and 0 for other-form repeaters) into the moderated regression equation.

Figures 3 and 4 present the regression lines for the new-form group and reference-form group, respectively. As shown, the pattern of regression lines and statistical results for their parameters were very similar in both test-form groups. The intercepts for the two repeater subgroups were statistically significantly different, but the slopes were not. The pattern of results indicates that equal performance on the anchor test predicts lower overall performance on the nonanchor items for those who have previously seen the anchor than for those who have not. In other words, the reference-repeater group performs better on the anchor without a commensurate improvement in performance on other items. This finding is important insofar as it indicates the need to differentiate repeaters based on the particular form of test that they took previously.

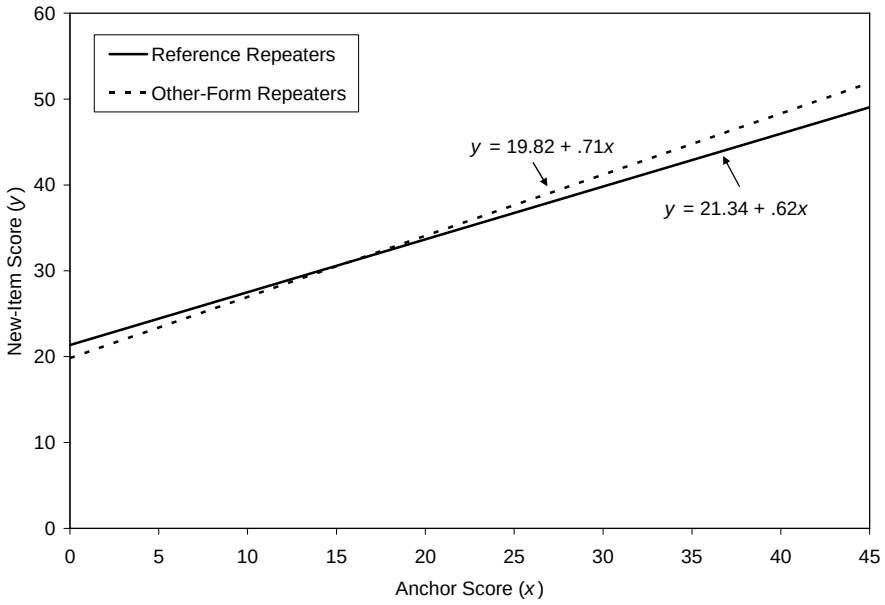

Figure 3. **Separate regression lines for reference repeaters and other-form repeaters in the new form (*X*) group.**

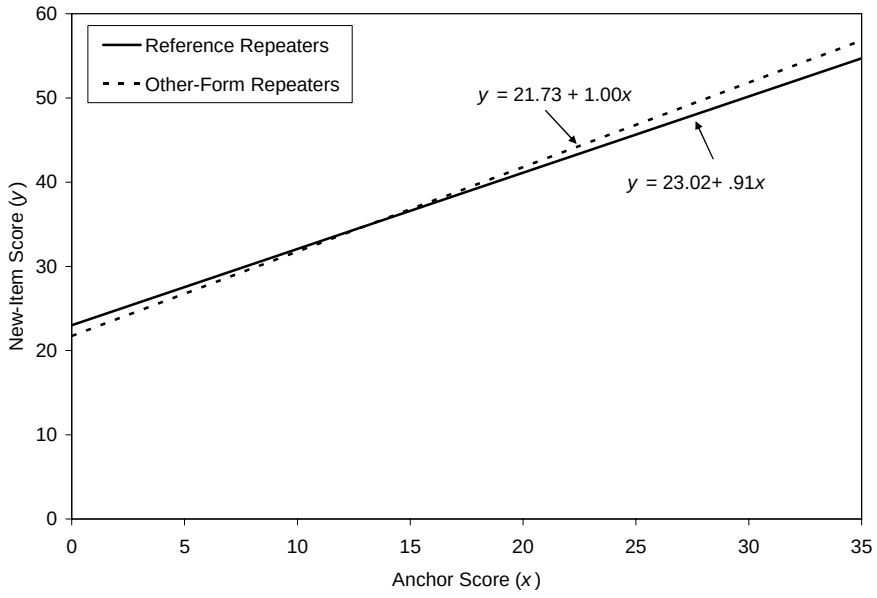

Figure 4. **Separate regression lines for reference repeaters and other-form repeaters in the reference form (*Y*) group.**

Subgroup Equating Analysis

Figure 5 presents the chained equipercentile equating functions, derived using the total group and various subgroups, at the new-form raw-score points where most examinees (about 98%) in the total group were located. The equating functions derived from total repeaters, other-form repeaters, and first-timers were very similar to the equating function derived from the total, indicating negligible differences among them. However, the equating function derived using the reference-repeater subgroup differed from the equating functions derived using the other groups, leading to differences in equated raw scores.

Figure 6 depicts equated raw-score differences between each subgroup equating and total-group equating, along with the DTM criterion (denoted by dashed lines). The solid line at zero denotes the total-group equating. The differences between the total-group and subgroup equatings fell within 0.5 raw-score units for the total-repeater and other-form-repeater subgroups. The difference for the first-timer subgroup was within the DTM range for raw scores higher than 58 but fell outside the range for raw scores lower than 59 This can be explained by the fact that the total equating function, particularly for the low-score region, was determined mainly by the repeaters, who tend to perform poorly, rather than the first-timers (see Figure 1).

As expected, the equating function derived using the reference-repeater group substantially differed from the equating function derived using the total group, yielding large positive differences for all score points. Those positive differences indicate that the new form was much more difficult than the reference form for the reference-repeater group than for the total group. Due to item exposure or practice, reference repeaters performed better on common items than on new (noncommon) items and therefore appeared more able than they would have if they had not previously seen the common items. Because reference repeaters did not do as well on the new (noncommmon) items, the new form appeared more difficult in the process of score equating. The differences were extremely large for raw scores higher than 80. However, as shown in Figure 1, almost no reference repeaters attained those high raw scores. These extreme differences were therefore somewhat artificial, caused by lack of data.

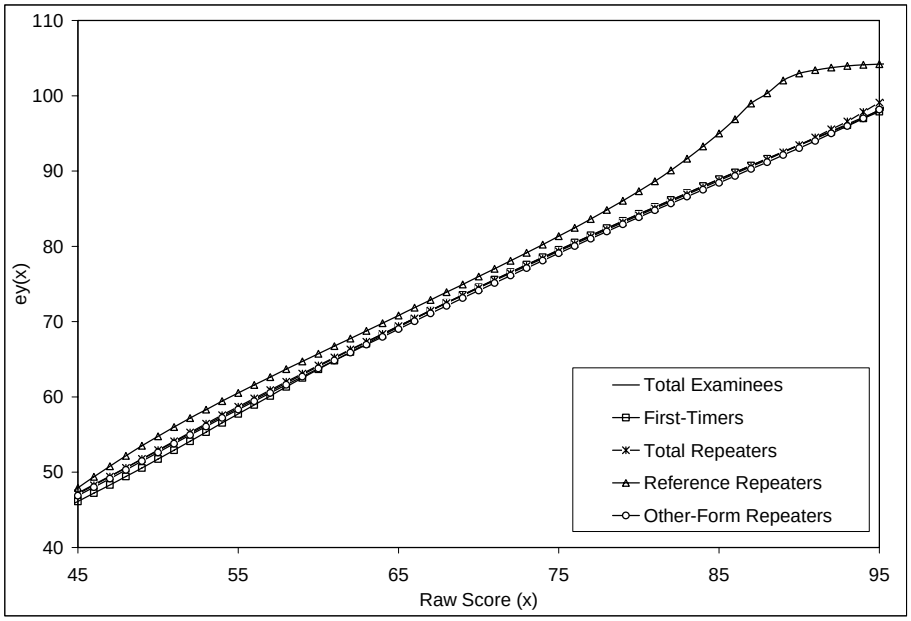

Figure 5. **Equating functions derived using total and various subgroup equating samples.**

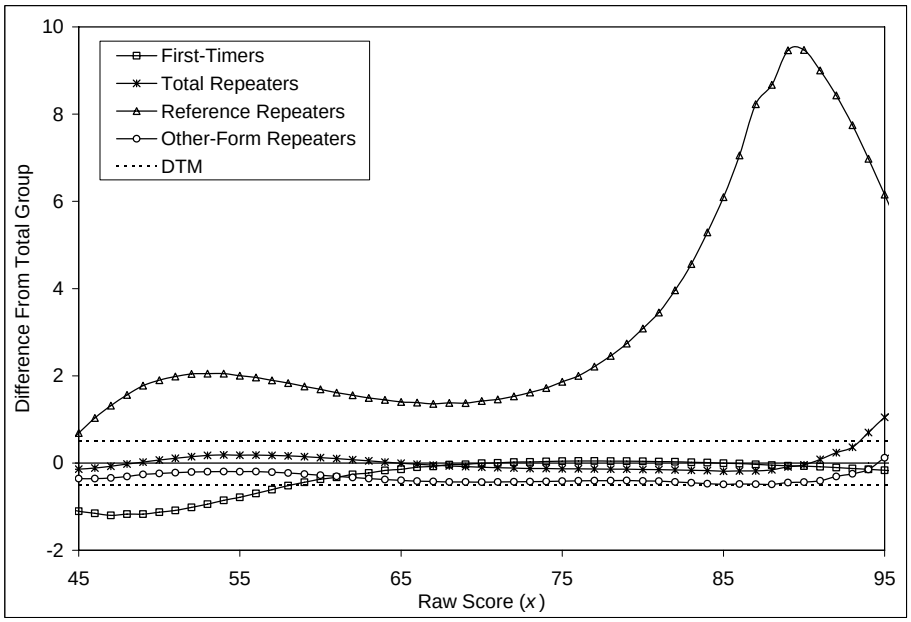

Figure 6. **Difference curves between the total-group equating function and subgroup equating function.**

Figure 7 shows how RMSD, depicted along with the DTM and REMSD measures, varies across score levels for the new form when the total group is divided into two subgroups: repeaters and first-timers. Both equal weights (e.g., 0.5 for each group) as well as unequal weights (proportional to subgroup size) were applied to the two subgroups in order to calculate the RMSD measure, but the difference between them was negligible across the score range. The solid line is the REMSD value of 0.24, which was almost constant regardless of the weighting methods. The dotted line at 0.5 denotes the DTM. The REMSD measure was about one half of the DTM, indicating that the equating function for each subgroup was sufficiently close to that of the total group. The RMSD values were much smaller than the DTM for most raw-score points, including the cut-score range (raw scores of 63–72). However, the RMSD was larger than the DTM at the lowest and highest scores. Given the closeness of all three equating functions (Figure 6) and the practically negligible RMSD and REMSD values for most raw-score points, the resulting equating function for this form can be considered invariant for both subgroups. The RESD values, which were calculated separately for each subgroup to assess subgroup dependence, also were practically negligible: 0.20 for the repeater subgroup and 0.27 for the first-timer subgroup. The *ew*REMSD values (0.11), summarized over the cut-score region, were almost negligible under the two-subgroup condition.

Figure 8 shows the same comparison as Figure 7 but with repeaters classified into those who previously took the reference form and those who took other forms. With the repeaters thus differentiated, the patterns of RMSD and REMSD dramatically changed, and the impact of weights was clearly noticeable. When the subgroups were weighted by their actual proportions in the total groups, the REMSD summary measure was twice as large as the DTM, and the RMSD values were larger than the DTM at the scale's low and high ends. However, the RMSD values were smaller than the DTM for raw scores in the 62–74 range, which includes the cut-score range of 63–72. This trend was more salient when the three subgroups were considered equally important regardless of their actual proportions in the total group. The RMSD values were much larger than the DTM for the entire score range, and the REMSD value (2.22) was four times as large as the DTM, clearly indicating subgroup dependence of equating functions. The extremely large RMSD values for the upper raw score range, however, is likely an artifact of the lack of

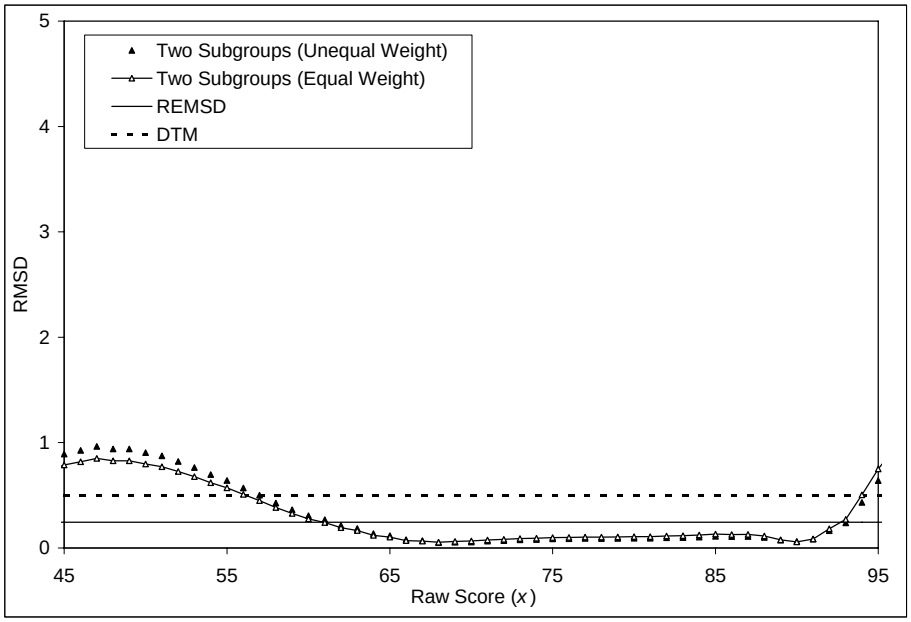

Figure 7. **Score-level root-mean-square-difference (RMSD) and overall root-expected-mean-square difference (REMSD) derived from comparing the total-group equating function to the two subgroup equating functions.**

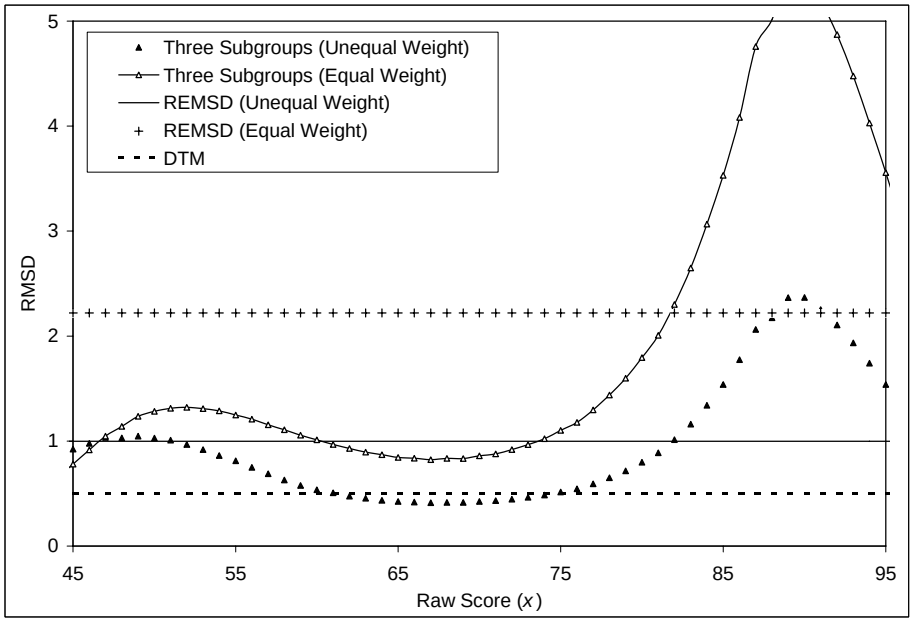

Figure 8. **Score-level root-mean-square difference (RMSD) and overall root-expected-mean-square difference (REMSD) derived from comparing the total-group equating function to the three subgroup equating functions.**

data (see Figure 1). Under the three-subgroup condition, the RESD values were 3.81 for reference repeaters and 0.41 for other-form repeaters. The *ew*REMSD value (0.86), summarized over the cut-score region, was much larger than the DTM under the condition in which the three subgroups were equally weighted. Again, such large RESD and *ew*REMSD values for reference repeaters indicate a substantial linking difference between the subgroup and the total group, which leads to different reported scores for the same examinees depending on the groups used for equating.

The results presented above were based on a single data set. The bootstrap resampling technique was applied to assess the variability of the RMSD values across 2,000 samples. Figures 9 to 12 show the 95% CIs for the RMSD values. Figure 9 presents the 95% CI of RMSD with undifferentiated repeater membership and both repeaters and first-timers weighted by their proportions in the total group. The 95% CI fell below the DTM in the denser portion of the score distribution (raw scores 66–80), which partially covered the cut-score range. However, the 95% CI became wider at the ends of the score range and covered the DTM. This result indicates that sample sizes in the distribution's lower and upper parts are insufficient to allow the conclusion that the population RMSD values exceed the DTM in these regions, although in this sample they appear to. Figure 10 presents the same information under the equal-weight condition. The pattern of results is almost identical, regardless of the weights imposed on the two subgroups.

Figure 11 presents the 95% CI for RMSD with reference-repeater and other-form-repeater subgroups differentiated and each subgroup weighted by its proportion in the total group. Although the 95% CI band narrowed in the cut-score range, it was much wider for raw scores over 80, indicating substantial fluctuation across samples. As shown in Figure 12, the fluctuation was much more pronounced when all subgroups were equally weighted. The DTM values fell outside of the 95% CI band across nearly all data points and were just within the band for raw scores from 61 to 73. Results indicated subgroup dependence of equating functions when subgroups were defined by the form of test previously taken.

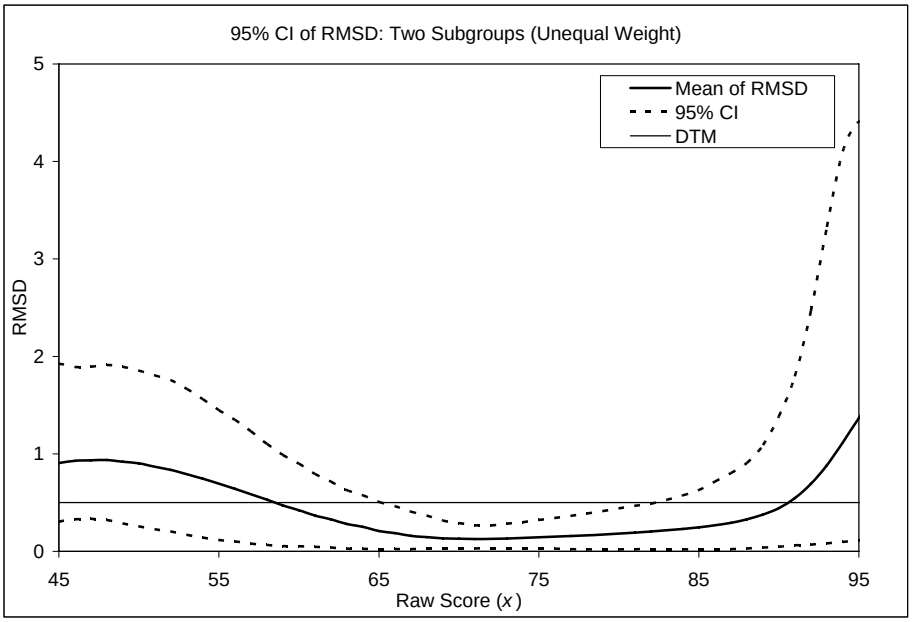

Figure 9. **The 95% confidence-interval (CI) band of root-mean-square difference (RMSD) with two subgroups weighted unequally based on their proportion of the total group.**

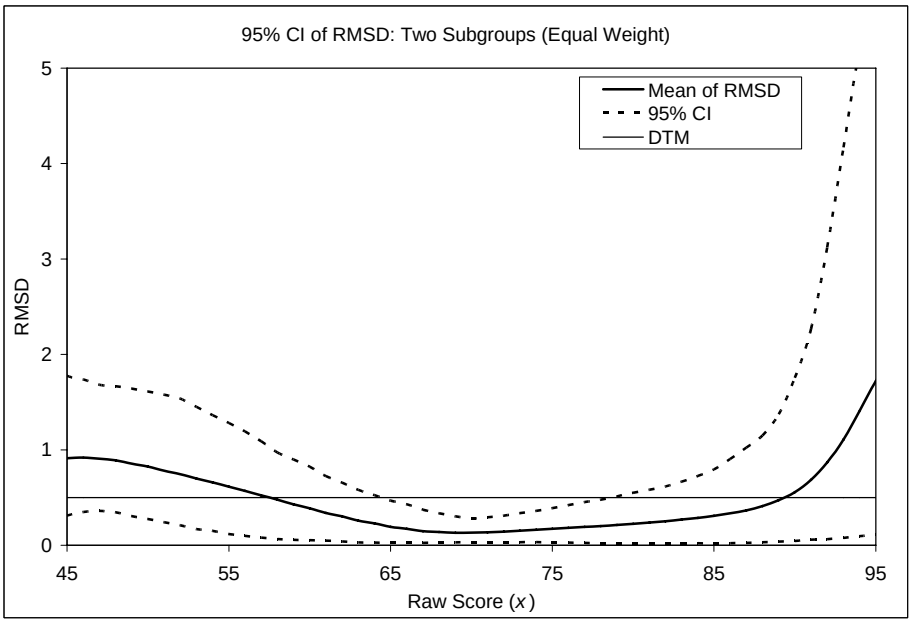

Figure 10. **The 95% confidence-interval (CI) band of root-mean-square difference (RMSD) with two subgroups weighted equally.**

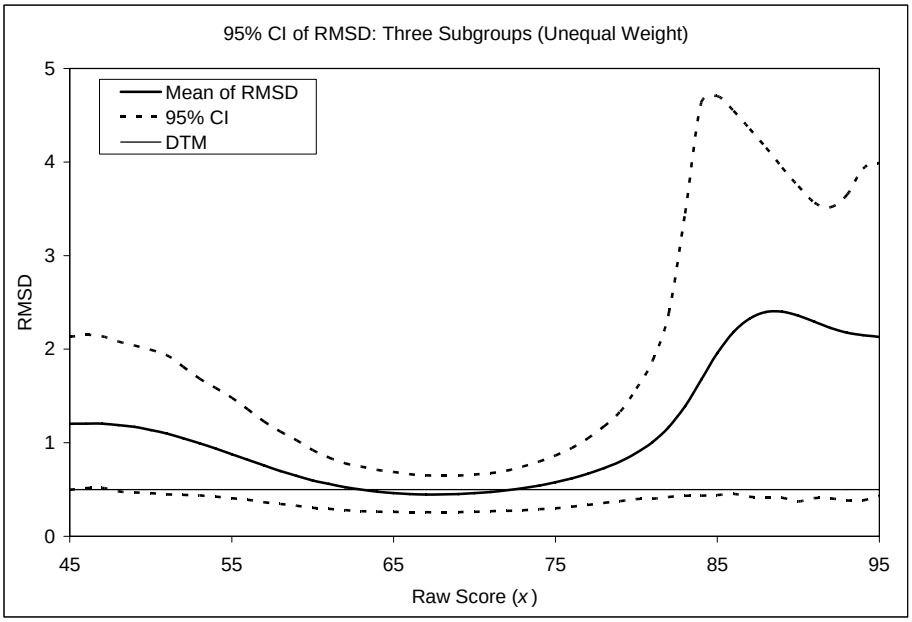

Figure 11. **The 95% confidence-interval (CI) band of root-mean-square difference (RMSD) with three subgroups weighted unequally based on their proportion of the total group.**

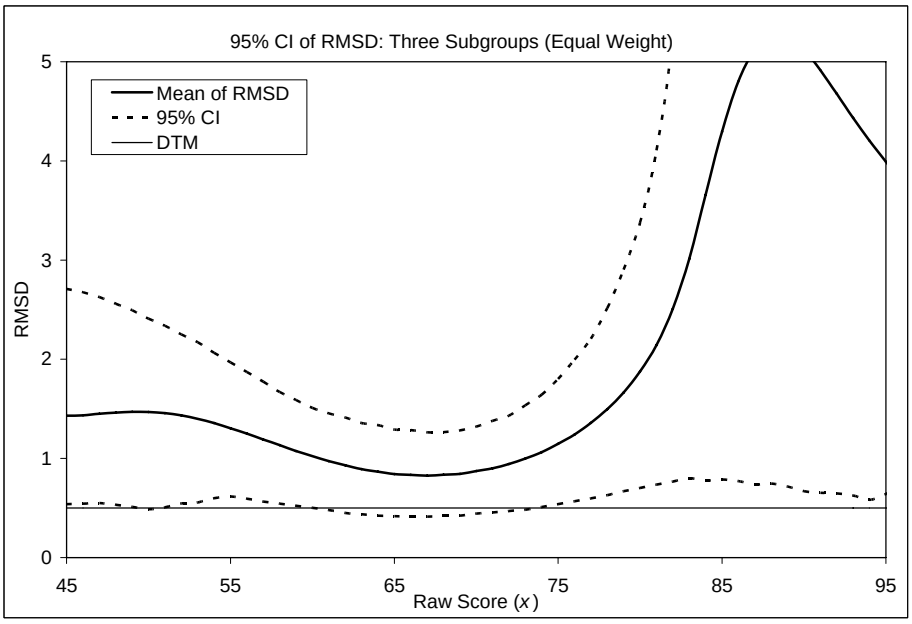

Figure 12. **The 95% confidence-interval (CI) band of root-mean-square difference (RMSD) with three subgroups weighted equally.**

Discussion

Dorans and Holland (2000) specified five requirements for equating that are often regarded as basic to all test equatings under consideration. Of these requirements, subgroup invariance is most critical for score equity assessment (SEA). SEA uses population invariance to assess the sensitivity of a statistical relationship between test scores and group membership (Dorans, 2004). To achieve score equatability, the resulting equated scores should have the same meaning no matter when or to whom the test was administered. Lack of subgroup invariance in an equating function indicates that the differential difficulty of the two forms is not consistent across those subgroups.

SEA indicates whether the linking relationship between two tests intended to measure the same construct remains constant across different subgroups. Under proper assessment conditions, score equatability is maintained. When tests are assembled using a well-established set of specifications, the relative difficulty of a particular version of a test changes as a function of score level in the same manner across subgroups; thus, the versions are related to one another in the same way across the subgroups. If the relative difficulty of different forms interacts with group membership or an interaction emerges among score level, difficulty, and group, subgroup invariance is not achieved. Checking SEA via subgroup invariance of equating functions can serve as a quality-control check to ensure that well-developed test assemblies remain within admissible tolerance levels with respect to equatability (Deming, 1982).

The present study investigated the sensitivity of equating to subgroups based on repeater status. This study examined the conditional RMSD and overall REMSD measures to assess score equatability as well as subgroup-specific RESD measures. The overall REMSD was smaller than the DTM, indicating negligible equating differences between the subgroups and the total group when repeaters were not differentiated according to their prior exposure to the reference form. Under this condition the curvilinear transformation was maintained reasonably well across repeaters and first-timers.

However, when we redefined the repeaters with respect to their previous experience with the common items, subgroup independence did not hold for the equating function. The equating function derived with repeaters who previously saw the reference form substantially differed from that derived with the total sample of examinees. The RESD value was very large when the equating function derived using the reference-repeater subgroup was compared to the total-group

equating function. The *ew*REMSD, summarized across the cut-score region, was much larger than the DTM, mainly due to the reference-repeater subgroup.

In most SEA analyses, evidence of population dependence suggests a need to reevaluate test-assembly specifications or linking methods. The results of this study, however, suggest a different remedy. In general, comparison of repeaters and first-timers revealed no differences between the equating functions across subgroups. Thus, familiarity with the material or the test format did not impact the equating function. Only when those individuals with previous exposure to the common items on the test form were examined did differences in equating functions emerge. Regression analyses, coupled with equating results, indicated that prior exposure to the common items changed the items' statistical characteristics, thereby invalidating them as anchor items. Here the indicated problem is neither the test specifications nor the linking method but a group of individuals with prior knowledge of the test items.

As shown in this study, repeaters' previous exposure to common items tends to pull the equating function away from the total-group equating function. Because the repeaters show enhanced performance on the common items but not on the rest of the test, the new-form sample appears more able than the reference-form sample, and the new form appears more difficult. This tendency increases as the proportion of repeaters in the equating sample increases. Such a situation would be particularly problematic with regard to licensure tests because the net effect would be a continued lowering of the raw scores that map to the qualification standard. Clearly, such repeaters should be excluded from the equating sample.

As described above, the repeater issue often is discussed in terms of exposure to common items. The assumption that repeaters perform better on all anchor items due to their previous exposure, though, seems unrealistic. Some items might be easier to memorize than others. After adjustments for differences in ability among subgroups, it might be useful to perform some statistical checks to discover which anchor items function differently across repeater and first-timer subgroups. Excluding problematic items from the anchor set might be an alternative way to reduce statistical equating error rather than excluding repeaters from the equating sample. If a large proportion of common items function differently across the two subgroups, the removal of many common items will bias the final equating function. In this situation, excluding repeaters from the equating sample might be the better choice.

This study has practical implications, but it also has limitations. First, the sample sizes for the subgroups were highly unbalanced; the proportion of reference repeaters was particularly small. Consequently, some extremely large RMSD, REMSD, and RESD values emerged, partly due to sampling variability. In this application we used the bootstrap procedure to compute the standard error of RMSD values and determine whether the differences obtained with the original data resulted from random error. However, implementing this procedure as a standard operation may be impractical for testing programs with strict deadlines. Second, in an operational setting, the examinees themselves report repeater information and other demographic variables. They may provide inaccurate information, and verifying their information is tedious.

For score equatability, equating must operate independently of the examinee subgroups from whom the data are gathered to develop the conversions. If the equating functions derived using different subgroups systematically differ, the interchangeability of test forms is questionable. Subpopulation invariance might be valid in some situations but not in others because this property is test-specific; it depends on the definition of subgroups and the characteristics of the ability being measured. Invariance of a certain subgroup on a certain test cannot be generalized to other subgroups or other tests. This means that invariance investigations should be conducted periodically with major subgroups for each test to ensure the fairness of score reporting, particularly for large-volume tests that affect high-stakes decisions.

References

Andrulis, R. S., Starr, L. M., & Furst, L. M. (1978). The effect of repeaters on test equating. *Educational and Psychological Measurement, 38,* 341–349.

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.

Cope, R. T. (1985, April). *Use versus nonuse of repeater examinees in common item linear equating with nonrandom groups*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Deming, W. E. (1982). *Out of the crises*. Cambridge: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41,* 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 equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281–306.

Holland. P. W. (2003). Overview of population invariance of test equating and linking. 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. 1–18). Princeton, NJ: ETS.

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

Liu, J., Feigenbaum, M., & Dorans, N. J. (2005). *Invariance of linkings of the revised 2005 SAT Reasoning Test to the SAT I: Reasoning Test across gender groups* (College Board Research Report No. 2005-6). New York: College Board.

Puhan, G. (2008). *What effect does the inclusion or exclusion of repeaters have on test equating?* (ETS Research Rep. No. RR-09-19). Princeton, NJ: ETS.

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

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

Notes

[1] It is possible that the test items, once exposed, could be compromised (e.g., by being posted on the Internet or otherwise widely distributed). In that case, both first-timers and repeaters who had not previously taken the anchor items may perform better on the anchor items than on the unique items. However, we do no consider this scenario here.

[2] In this situation, preserving only univariate moments yields a slightly better fit of the marginal distributions than is the case when the first bivariate moment also is preserved. Such a strategy is possible because chained equating operates only on the margins. In any event, differences in equating results with and without preservation of the bivariate moment are negligible.

[3] Equal weight is the inverse of the total number of scores and thus the weight is 0.1 in this case.

[4] Not all states use the same cut scores.