Research Report

Evaluating Subpopulation Invariance of Linking Functions to Determine the Anchor Composition for a Mixed-Format Test

Sooyeon Kim and Michael E. Walker

December 2009

ETS RR-09-36

Evaluating Subpopulation Invariance of Linking Functions to Determine the Anchor Composition for a Mixed-Format Test

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

December 2009

Abstract

We examined the appropriateness of the anchor composition in a mixed-format test, which includes both multiple-choice (MC) and constructed-response (CR) items, using subpopulation invariance indices. We derived linking functions in the nonequivalent groups with anchor test (NEAT) design using two types of anchor sets: (a) MC only and (b) a mix of MC and CR. In each anchor condition, we derived the linking functions separately for males and females and then compared those subpopulation functions to the total group function. In the MC-only condition, the difference between the subpopulation functions and the total group function was not trivial in a score region that included cut scores, leading to inconsistent pass/fail decisions for low-performing examinees in particular. Overall, the mixed anchor was a better choice than the MC-only anchor to achieve subpopulation invariance between males and females. The research reinforces subpopulation invariance indices as a means of determining the adequacy of the anchor.

Key words: Mixed-format test, NEAT design, gender differences, subpopulation invariance, equating

Acknowledgments

We would like to thank Mary C. Grant, Wen-Ling Yang, and Daniel Eignor for their many helpful comments on the earlier draft of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer and Ruth Greenwood.

Background

Mixed-Format Tests

Many large-scale testing programs include both constructed-response (CR) and multiple-choice (MC) items in their assessments. MC items are economically practical and ensure objective and reliable scoring. CR items may be difficult to score objectively and reliably, but they may measure examinees' understanding of particular content more deeply than do MC items. Because both MC and CR items have strengths as well as weaknesses, many assessments tend to be mixed format, including both types of items.

As with other standardized tests, these mixed-format tests must be equated to ensure equivalence of scores across test forms. Most often, equating occurs in the context of the nonequivalent groups with anchor test (NEAT) design, in which a set of items common to both the new and old forms is used to place both forms on the same scale. These common items should represent the entire test form in content and difficulty.

NEAT equating has proven difficult with mixed-format tests: Identification of a satisfactory anchor test has been a particular problem in equating tests with a CR component. In many cases, for example, CR items are not re-used across different test forms because memorizing and sharing them is easy (Muraki, Hombo, & Lee, 2000); thus, no common CR items are available for equating. Even if the same CR items are used, raters' standards in scoring those items tend to change across administrations. Some practitioners have suggested using MC items as anchors to control for differences among test forms that include CR items (e.g., Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). Several empirical studies suggest, however, that use of an all-MC anchor produces biased linking results (Kim & Kolen, 2006; Kim, Walker, & McHale, in press; Li, Lissitz, & Yang, 1999), possibly because MC and CR items measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002). Anchors consisting of MC items alone may not represent the entire test content and thus may not produce satisfactory linkings.

Various linking designs and methods have been discussed thoroughly in the literature (Kolen & Brennan, 2004). These can be applied readily to the linking of tests with MC items. Several IRT linking methods have also been adapted for use with tests that include CR items (Baker, 1992; Cohen & Kim, 1998; Kim & Kolen, 2006). Nevertheless, applying these methods to tests including CR items can still be problematic. CR items take longer to answer than do MC items; therefore, a test with CR items will necessarily contain fewer items and often be less reliable than will a test with all MC items administered within the same time limit. Issues such as test length and reliability directly impact linking quality (Fitzpatrick & Yen, 2001).

The Use of Population Invariance Indices

To ensure fairness to examinees who take different forms, the scores on the forms should be equated to make them equivalent to one another. Once properly equated, the resulting scores should have the same meaning no matter which version of the test was administered, or when, or to whom. This type of score equity is difficult to test directly. Instead, practitioners generally assess whether equating has been successful by examining the population invariance of the linking function. Theoretically, the population invariance requirement means that the equating function must operate independently of subpopulations of examinees from whom the data are drawn to develop the conversion (Angoff, 1971). This requirement is necessary for equating to take place. If the function relating two test forms is not invariant across subpopulations, the new and reference test forms have not been equated and the interchangeability of the linked scores is questionable.

In reality, the particular population used to link two tests always affects the linking function, so that population invariance is never achieved absolutely. Instead, the question becomes whether population invariance holds closely enough such that the linking function is not differentially affected across subpopulations (Dorans & Holland, 2000). What characteristics of the data lead to subpopulation differences in linking functions? When tests are assembled using a well-established set of content and statistical specifications, the relative difficulties of different versions of a test will likely change as a function of score level in the same manner across subpopulations; thus, the versions are related to each other in the same way across the subpopulations. If the relative difficulties of different forms interact with group membership, or if an interaction emerges among score level, difficulty, and group, subpopulation invariance is not achieved. This issue could become more critical in score linking with mixed-format tests due to a potential group-by-item format interaction. The situation is further complicated in the NEAT design, in which the relationships between the anchor and each of the two test forms to be equated must remain constant across subpopulations. This condition would be more likely to hold when the composition of the anchor matches those of the test forms to be equated.

This research project examined the anchor compositions that would lead to equated test scores in the case of mixed-format tests. Ideally, the anchor would be a miniature version of the total test, containing both MC and CR items. For practical reasons, CR items are often not available for the anchor set, a situation that would appear to preclude equating in the strictest sense insofar as it would likely result in population dependence of the linking functions. As mentioned previously, however, population invariance is a matter of degree. The question is whether or not invariance holds closely enough to achieve a reasonable equating relationship. It may very well be that under some conditions it is not necessary to include CR items in the anchor to equate a mixed-format test. It would be useful to know when an MC-only anchor is sufficient to establish an equating relationship. One approach would be to examine the invariance of linking relationships across major subpopulations of interest when the MC-only anchor is used. Checking subpopulation invariance of linking functions could serve to determine which anchor composition performs better in achieving score equity in mixed-format tests. If linkings across subpopulations lead to the same relationship between old and new forms, use of the MC-only anchor would be supported.

Gender Effects on Mixed-Format Tests

The decision to use MC, CR, or both is not simply a choice of response formats; it is a choice of skills to measure. A common result of the difference in skills that MC and CR items measure is a group-by-format interaction. Several previous research studies indicate that when males and females perform equally well on MC, females do better on CR; conversely, when males and females perform equally well on CR, males do better on MC. Petersen and Livingston (1982) found such a difference when comparing male and female students' performance on the College Board's English Composition Test with Essay within each of four ethnic subpopulations. Other studies using large-scale data from Advanced Placement Program® (AP®) examinations (Mazzeo, Schmitt, & Bleistein, 1992; Willingham & Cole, 1997) showed that the strength of the gender-by-format effect varied across academic subjects. Later, Livingston and Rupp (2004) examined performance of male and female examinees on MC and CR tests for beginning teachers and found that the strength of these format gender differences varied as a function of subjects and school levels.

If such effects result in differences in the relative difficulty of test forms across gender groups, or in differences in the relationship between the anchor and the total test across groups, then the effects could seriously undermine the test linking process. Invariance of linking functions across groups is necessary for successful equating. Given the above arguments, gender subpopulations could play a particularly important role in the linking process for mixed-format tests.

Purpose

The major purpose of the present study was to use subpopulation invariance indices to examine the appropriateness of anchor composition in a large-scale mixed-format licensure test. Linking functions were derived in the NEAT design using two types of anchor sets: (a) MC only and (b) a mix of MC and CR. The subpopulation invariance of score linking was determined by comparing linking functions derived separately for male and female examinees to the linking function derived using the combined group. The study design and methodology allowed examination of group-by-item format interactions and their direct effects on the linking process. The study focused on classical linking methods to answer two major questions: (a) Does a gender-by-format interaction impact linking functions when the test contains CR items? and (b) what anchor test composition (MC and CR items, or MC items only) is more likely to achieve score equity?

Methods

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 24 rights-scored MC items and 12 CR items; 12 MC and 6 CR items[1] were common across new form *X* and old form *Y*. The CR items were scored by a single rater on a 0-2 scale and weighted by 2. Thus, the maximum total test score was 72, and examinees could earn twice as many raw-score points on the CR section as on the MC section. The scaled score of this test ranged from 100 to 200, rounded to the nearest multiple of 1.0.

Descriptive statistics are summarized in Table 1. The total new form group was more able than was the total old form group. Their mean scores on the mixed anchor differed by 0.84 correct answers, an effect size of 0.14, whereas their mean scores on the MC-only anchor differed by 0.12 correct answers, an effect size of 0.07. The difference in the mean scores on the two forms (an effect size of 0.27) therefore appears to be attributable mainly to the difference in difficulty between the two forms.

In both new and old form groups, female examinees were higher in ability as measured by the raw test scores than were male examinees. Although females performed better than males on both MC and CR, the gender difference was larger on the CR portion than on the MC portion of the test. As the data presented in Table 2 indicate, the trend was the same with the total MC and CR scores, both of which included common and non-common items. Figure 1 presents the relative frequency distributions of new-form scores in the two subpopulations, given as percentages of the total group. Figure 2 presents the relative frequency distributions of old-form scores in the two subpopulations.

Procedure

In the NEAT design, the linking function derived using either males or females was compared to that which was derived using all examinees to determine whether the resulting linking function from the total group would yield scores comparable to the subpopulation linking functions. The study included the following three steps:

1. Obtain total group and subpopulation linking functions using a mixed (MC and CR) anchor in the NEAT design. The linking relationship between the new (X) and old (Y) forms was derived using three groups: (a) total examinees, (b) males, and (c) females. Raw-to-scaled score conversions were obtained using the chained equipercentile linking method.[2]
2. Compare total group and subpopulation linking functions. The male subpopulation function was compared to the female subpopulation function. The linking function derived using each subpopulation was compared to the total group linking function. The differences were quantified across both subpopulations using the conditional root mean square difference (RMSD) and the summary root expected mean square difference (REMSD) deviance measures (see Equations 1 and 2 in the following section, *Deviance Measures*). In general, a negligible difference indicates that the linking relationship is not influenced significantly by the subpopulation used in deriving that function. In addition, the difference between each subpopulation linking function and the total-group linking function was separately quantified to assess more linking

Table 1

Summary Statistics for the New and Old Form Groups

Test				Total		Mixed anchor			MC-only anchor			Scaled score	
form	Group	*N*	(%)	*M*	*SD*	*M*	*SD*	*r*	*M*	*SD*	*r*	*M*	*SD*
New	Total	3,336	(100%)	46.68	9.59	21.47	5.87	.89	7.50	1.75	.58	170.87	11.87
form	Male	1,511	(45%)	44.75	9.40	20.53	5.85	.89	7.34	1.79	.57	168.52	11.53
(*X*)	Female	1,825	(55%)	48.27	9.45	22.25	5.77	.89	7.63	1.71	.58	172.82	11.80
Old	Total	2,248	(100%)	44.02	10.04	20.63	5.78	.89	7.38	1.76	.57	169.21	12.01
form	Male	987	(44%)	41.41	9.79	19.33	5.65	.88	7.15	1.81	.56	166.19	11.71
(*Y*)	Female	1,261	(56%)	46.06	9.77	21.64	5.67	.89	7.56	1.70	.57	171.58	11.71

Note. *r* indicates the correlation between total and anchor scores.

Table 2

Summary Statistics of Multiple-Choice and Constructed-Response Scores in the New and Old Form Groups

Test form	Group	*N*	(%)	MC		CR		
				M	*SD*	*M*	*SD*	*r*
New	Total	3,336	(100%)	16.45	3.30	30.23	7.69	.43
form	Male	1,511	(45%)	16.03	3.35	28.72	7.56	.40
(*X*)	Female	1,825	(55%)	16.79	3.22	31.49	7.58	.44
Old	Total	2,248	(100%)	16.39	3.69	27.63	7.86	.44
form	Male	987	(44%)	15.86	3.74	25.55	7.62	.42
(*Y*)	Female	1,261	(56%)	16.80	3.59	29.25	7.66	.43

Note. r indicates the correlation between MC and CR scores.

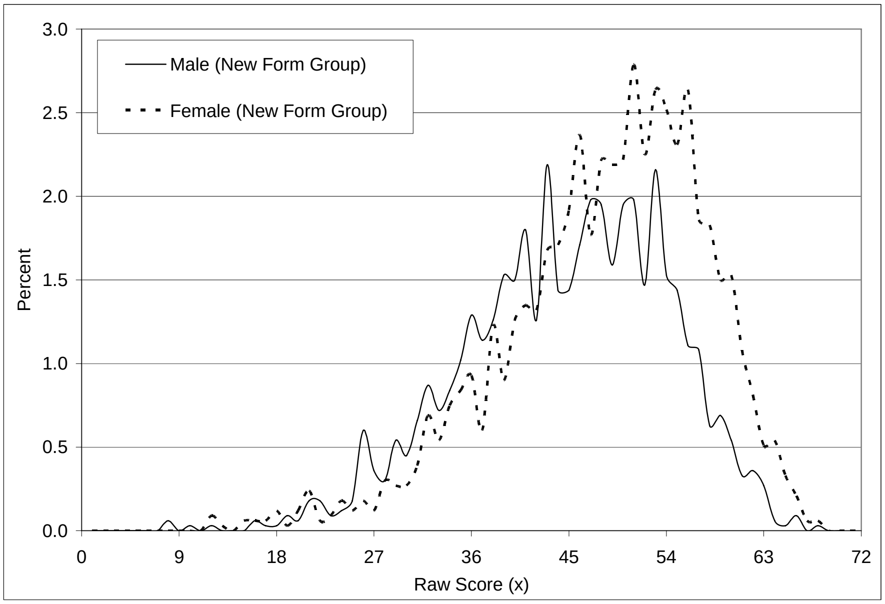

Figure 1. **Relative frequency distribution of new form *x* scores in the male and female subpopulations.**

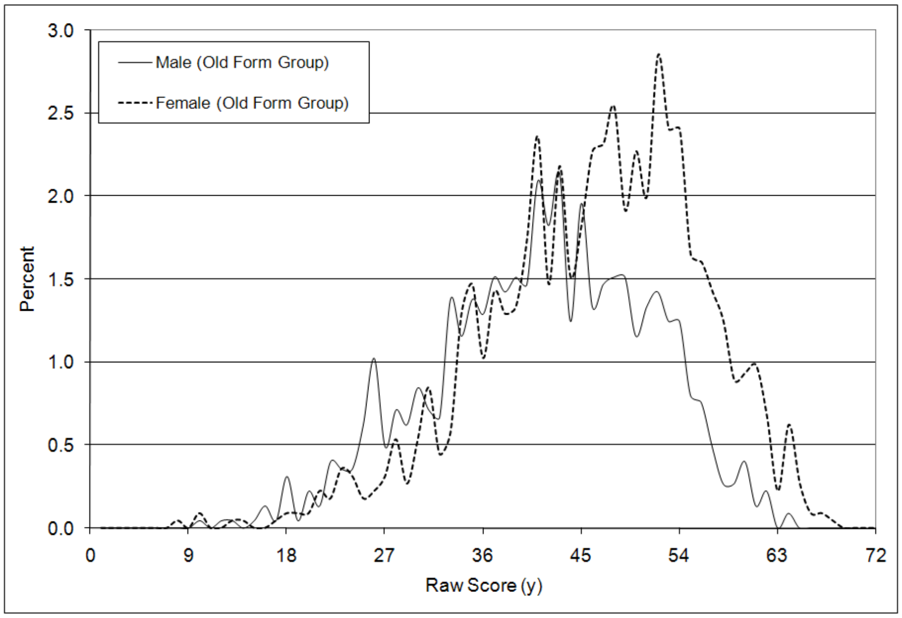

Figure 2. **Relative frequency distribution of old form *y* scores in the male and female subpopulations.**

transformation (root expected square difference, or RESD; see Equation 4). The impact of the three different conversions (i.e., those derived using male, female, and total samples) on examinees' pass/fail designations was also assessed.

3. Estimate precision of the deviance measures. A total of 2,000 replications were obtained for each linking function 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 old-form and new-form group until bootstrap samples included the same number of examinees as the actual old- and new-form groups. Both the old- and new-form samples then were divided into two mutually exclusive subpopulations, according to examinees' gender. In each replication, three groups (total group, male subpopulation, female subpopulation) were formed; for each group, we equated the new-form scores to the old-form scores using the chained equipercentile method and calculated the RMSD using proportional weights. The 95% confidence interval (CI) for the RMSD measure, which covers RMSDs in the 2.5th to 97.5th percentile range, was constructed on the basis of the 2,000 replications to evaluate linking differences from a statistical perspective.

Steps 1 to 3 were repeated to assess the MC-only anchor condition.

Deviance Measures

Four deviance indices were used as population invariance measures. 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 raw cut-score region (29 to 44),[4] to examine the impact of subpopulation on examinees' pass/fail designations. In addition, the RESD was computed as the weighted average of the squared differences between each subpopulation linking function and the total-group linking function at each raw score level. Thus, each subpopulation had a single summary RESD value in scaled score units. 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 various indices are defined as follows:

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

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

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

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

where x represents each raw score point, $e_{yij}(x)$ indicates the linking function in the *j*th subpopulation, $e_{yi}(x)$ represents the linking function in the total group, w_j denotes the proportion of subpopulation *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 subpopulation size was not seriously unbalanced. Across both groups, the proportion of females was 55% to 56% and the proportion of males was 44% to 45%. Thus, neither subpopulation heavily influenced the RMSD or REMSD measures. Therefore, only the proportional (i.e., unequal) weight derived from the actual relative size of each subpopulation was imposed on each subpopulation when calculating RMSD and REMSD.

To determine when the REMSD was large enough to warrant concern about form equatability, the notion of the score *difference that matters* (DTM; Dorans & Feigenbaum, 1994), defined as half of a scale score point in the raw-to-scaled score transformations, was used. Half a point was used here because we would expect any differences less than half a point to round to the same integer reported score value. Thus, we can expect differences less than half a point not to matter to the examinee.

Results

Regression Analysis

One supposition underlying this study was a potential gender by item format interaction effect, which can produce linking bias. As the data presented in Table 2 show, the correlation between the MC portion and the CR portion of the test was slightly stronger for females than for males; this pattern was consistent across the old- and new-form groups. To examine the interaction effect in more detail, moderated regression analyses were performed predicting CR total scores from MC total scores, gender (male vs. female), and the MC by gender interaction. The overall *R*-squared for the full fitted model was 0.20 for the new-form group and 0.23 for the old-form group. From this overall analysis, separate regression lines for the two subpopulations then were computed to predict CR total scores from the MC total score by substituting gender information (1 for females and 0 for males) into the moderated regression equation. These regressions are shown in Figures 3 and 4.

Figure 3 reveals a significant difference in the predicted conditional total CR score means across the male and female groups in the new form (X) group.

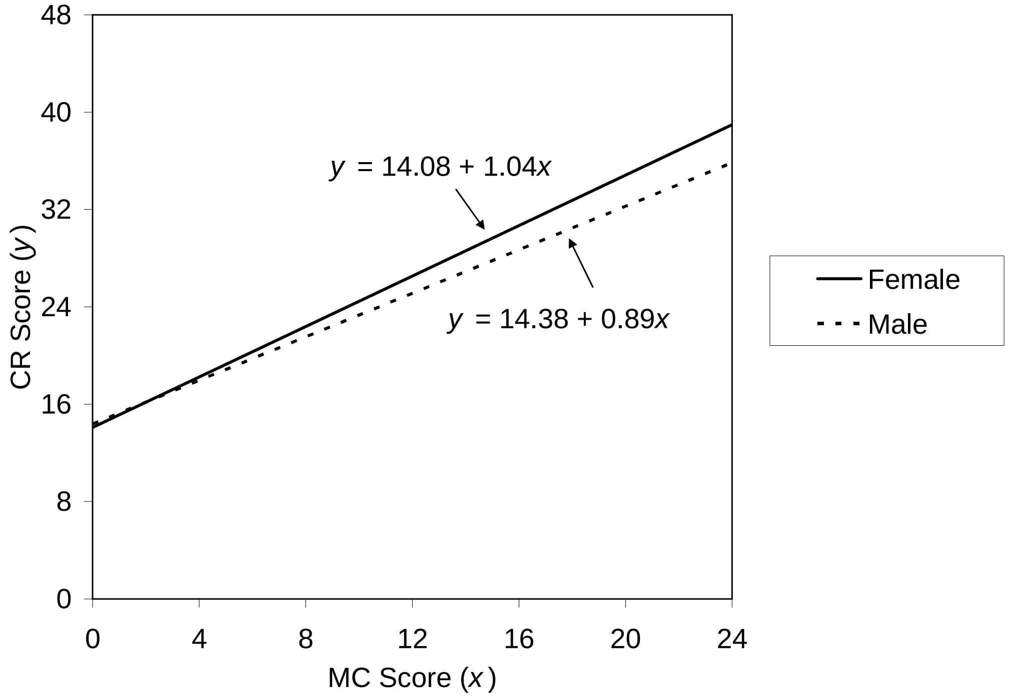

Figure 3. **Separate regression lines for males and females in the new form (*X*) group.**

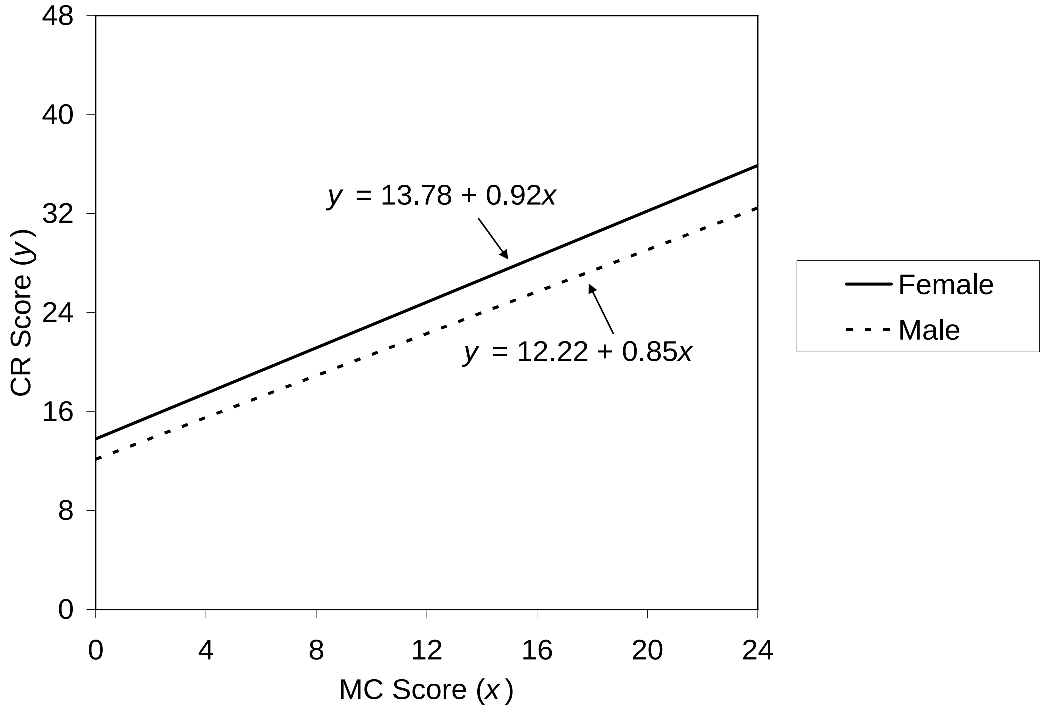

Figure 4. **Separate regression lines for males and females in the old form (*Y*) group.**

Figure 4 presents the same information in the old form (*Y*) group.

In both test form groups, equal performance on the MC items predicts lower overall performance for males than for females on the CR items. The difference was greater toward the higher than the lower end of the MC score range. This trend was more salient for the new-form group than for the old-form group, leading to a statistically significant MC by gender interaction ($p = 0.05$) in the new-form group. This finding is important insofar as it reveals an interaction effect that might result in linking bias for mixed-format tests as a function of anchor composition. More importantly, we would expect that any anchor that does not represent both the MC and CR sections proportionate to the total test would result in different linking functions across gender subpopulations because of the differential relationships between the components of the test across the two groups.

Subpopulation Linking Analysis

To determine if the potential differences in linking functions across subpopulations did manifest themselves when different anchor tests were used, the new form was linked to the old form for the total group and for the two subpopulations. Deviance measures among the resulting raw-to-scaled score conversions were compared to the DTM to assess how substantial the differences among the conversions were. Figure 5 presents the scaled score chained equipercentile linking functions in the mixed-anchor condition, derived using the total group and the two subpopulations, for all new-form raw score points above the first percentile. The two subpopulation functions departed from the total group function for raw scores lower than 32, but their departures were in opposite directions.

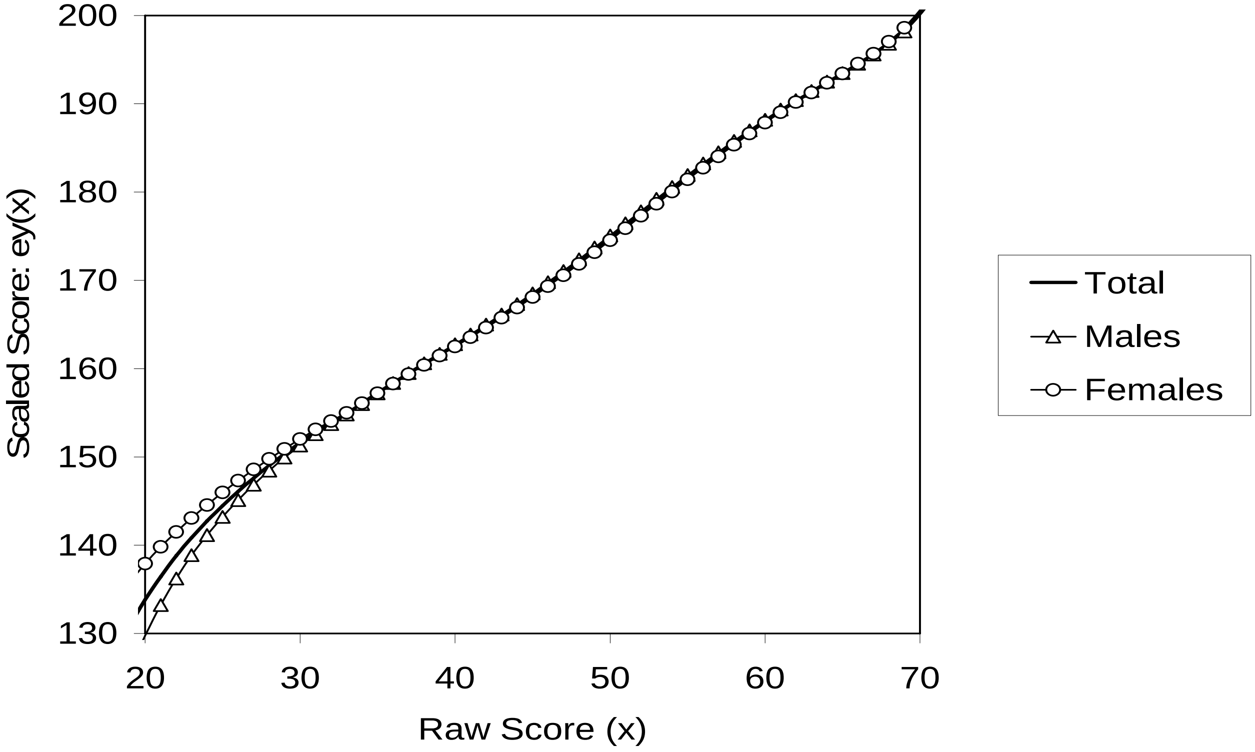

Figure 5. **Raw-to-scaled linking functions derived using total and two subpopulations in the mixed-anchor condition.**

Figure 6 presents the same comparison under the MC-only anchor condition. The departure of the two subpopulation functions from the total group function was not as marked as in the mixed-anchor condition, particularly at the lower scores.

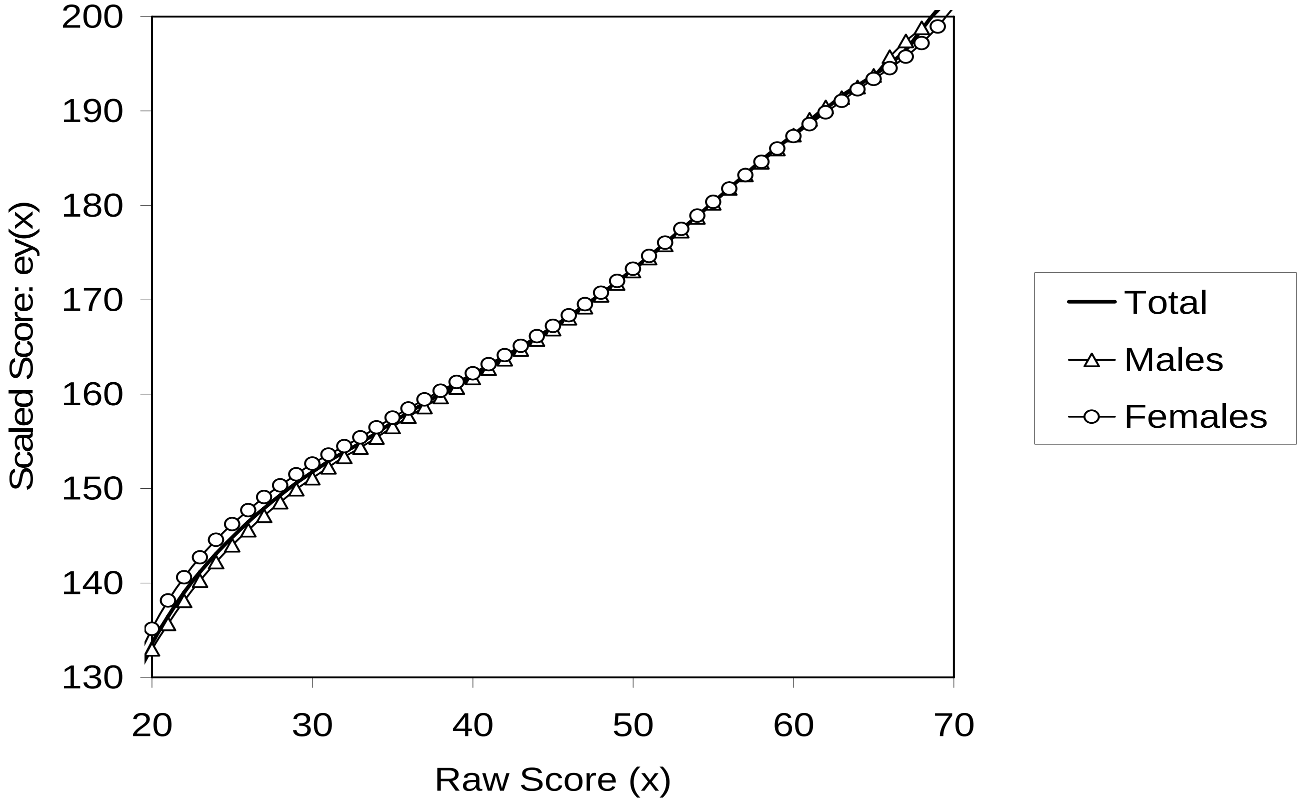

Figure 6. **Raw-to-scaled linking functions derived using total and two subpopulations in the MC-only anchor condition.**

Figure 7 depicts scaled score differences between each subpopulation equating function and the total group equating function in the mixed-anchor condition, along with the DTM criterion (denoted by dashed lines). The solid line at zero denotes the total group equating. The difference between the total group and each subpopulation equating fell within 0.5 scaled-score units for raw scores above the fifth percentile (raw score of 29), where all the cut scores were located. The differences for both subpopulations fell outside the DTM range for raw scores lower than 29 and their departures were in opposite directions. For the low performing examinees, the total group conversion underestimated the new form difficulty for females and overestimated the new form difficulty for males.

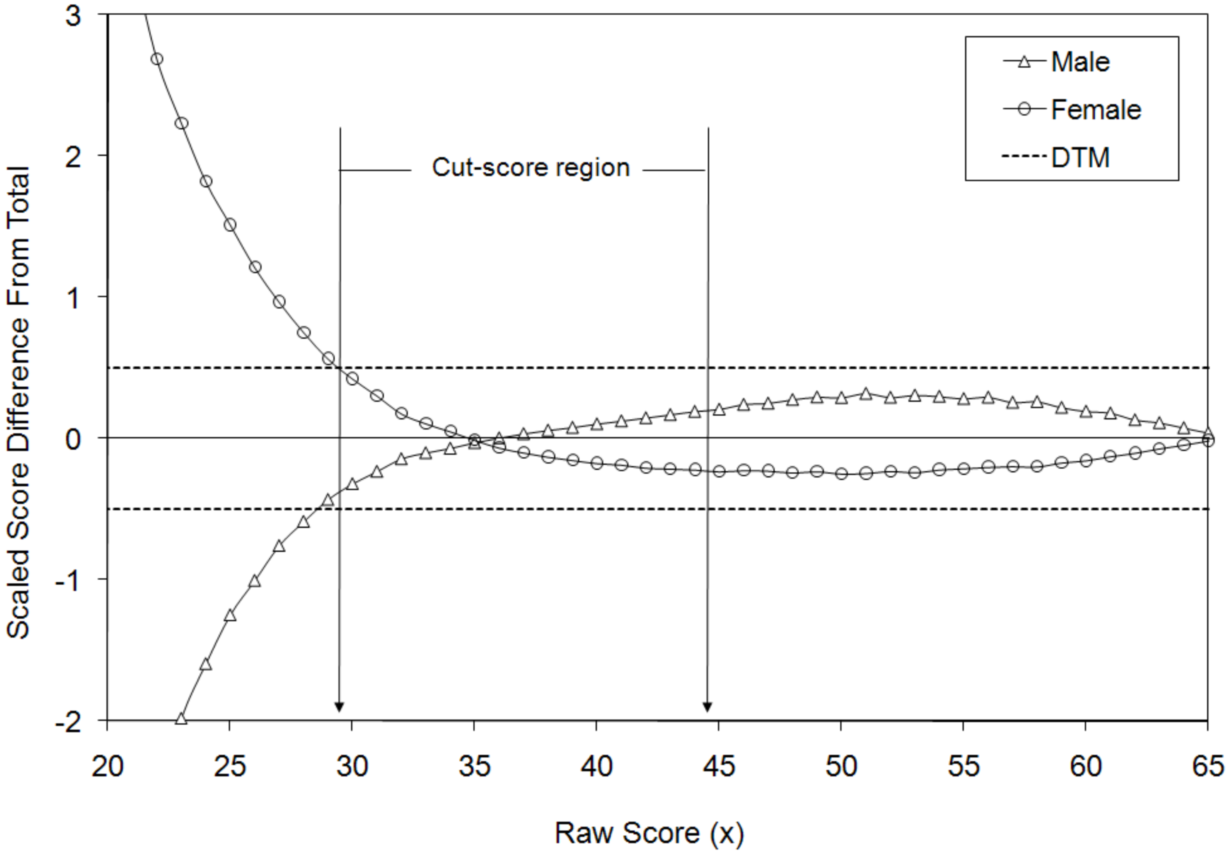

Figure 7. **Scaled-score difference curves between subpopulation equating function and the total group equating function in the mixed-anchor condition.**

Figure 8 presents the same comparison as Figure 7 under the MC-only anchor condition. The difference between the total group and each subpopulation equating fell within 0.5 scaled-score units for raw scores above the tenth percentile (raw score of 34), where many cut scores were located. The differences for both subpopulations fell outside the DTM range for raw scores lower than 34, which included a few cut scores. The difference between each subpopulation function and total group function, however, was smaller in the MC-only condition than in the mixed-anchor condition for raw scores lower than 27 and for raw scores from 44 to 60, where many examinees' scores fell.

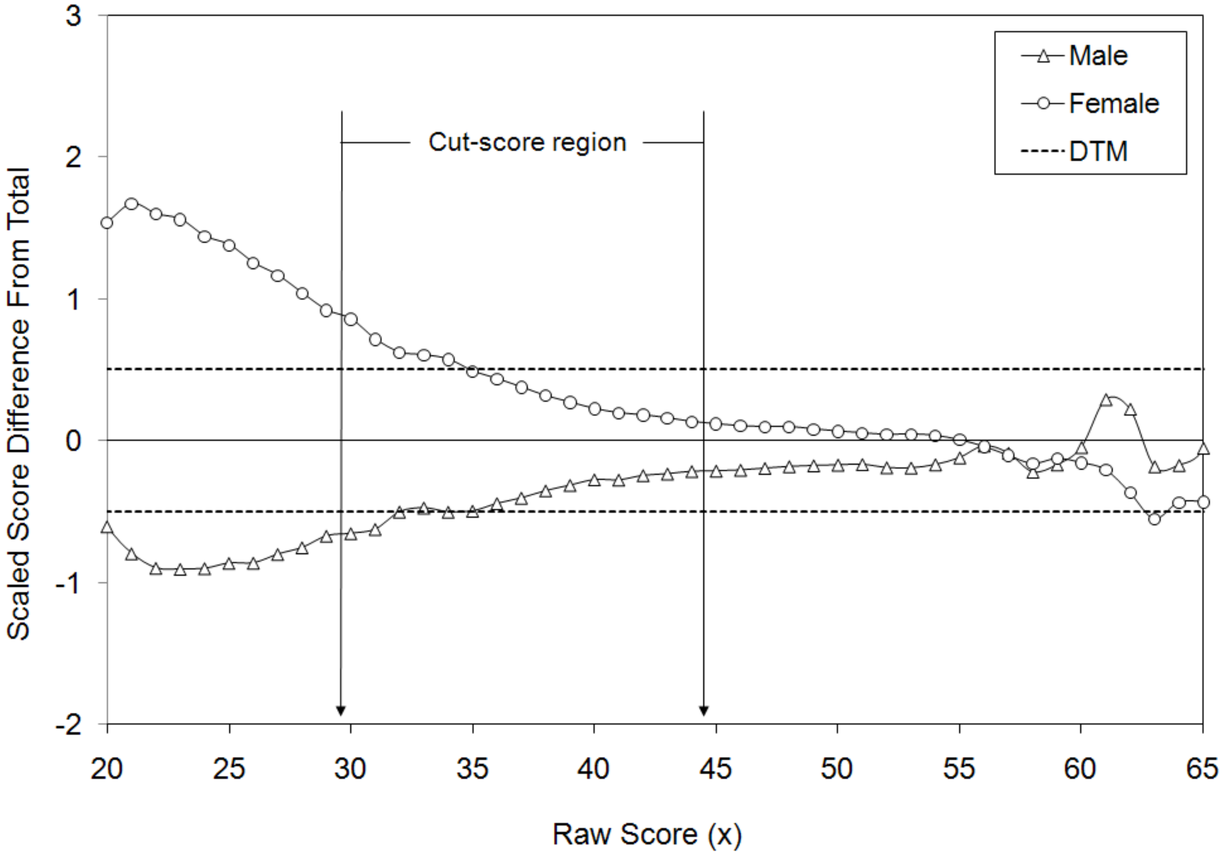

Figure 8. **Scaled-score difference curves between subpopulation equating function and the total group equating function in the MC-only anchor condition**.

Figure 9 presents the conditional scaled-score unit RMSD, along with REMSD and DTM, in the two anchor conditions. In the mixed-anchor condition, the linking functions derived from male and female subpopulations were similar to that derived from the total for raw scores higher than 29, including the cut-score region. In the MC-only condition, however, the linking functions derived from male and female subpopulations differed more substantially from the total group linking function within the cut-score region, thus leading to potential differences in reported scores. The summary REMSD value, however, was smaller in the MC-only anchor condition (0.37) than in the mixed-anchor condition (0.64). Such high REMSD for the mixed-anchor condition was caused mainly by the extremely large RMSD values for the low end of the score range. The *ew*REMSD for the cut-score region was 0.26 with the mixed anchor and 0.52 with the MC-only anchor. Under the MC-only anchor condition, the RESD values were 0.35 for males and 0.39 for females. Under the mixed-anchor condition, these values were 0.59 for males and 0.69 for females, which were larger than the DTM.

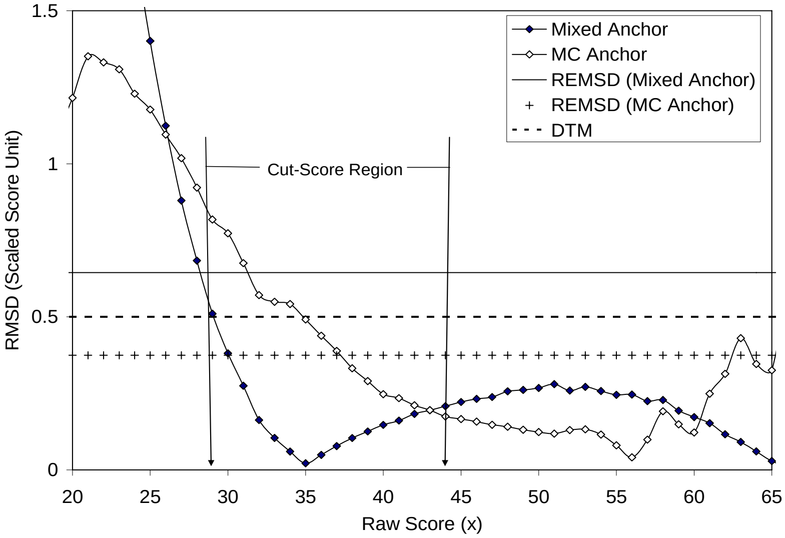

***Figure 9.* Score-level RMSD and overall REMSD derived from comparing the total-group linking function to the two subpopulation linking functions.**

Bootstrapped Resampling

The results presented thus far were based on a single data set. The bootstrap resampling technique was applied to assess the variability of the RMSD values. We compared total and subpopulation linking functions using the 2,000 bootstrapped samples in the two anchor conditions and constructed 95% CIs for the RMSD values. These CIs were useful in assessing the importance of the sample deviance values obtained in this study.

Figures 10 and 11 present the 95% CI of scaled-score unit RMSD in the mixed-anchor and MC-only anchor conditions, respectively. In both conditions, the 95% CI did not fall below the DTM even in the dense portion of the score distribution. The 95% CI became wider at the ends of the score range and covered the DTM. In the mixed-anchor condition, however, the DTM values fell outside the 95% CI band for scores below 21. As shown in Figure 11, the fluctuation was much more pronounced in the MC-only anchor condition, indicating substantial fluctuation across samples and a concomitant lower level of precision. The result for the MC-only anchor condition indicates that samples in the cut-score region are too small to support a definitive conclusion that the population RMSDs exceed the DTM in that region, although in this sample they appear to do so.

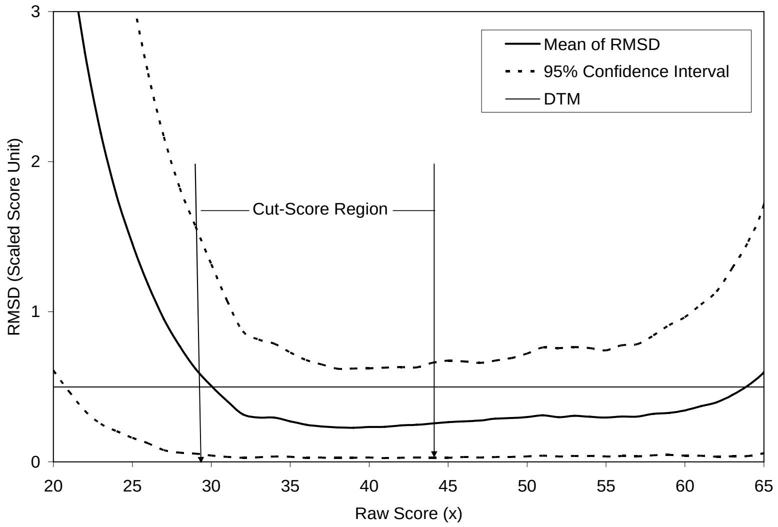

***Figure 10.* The 95% confidence interval (CI) band for the scaled-score unit RMSD: Mixed-anchor condition.**

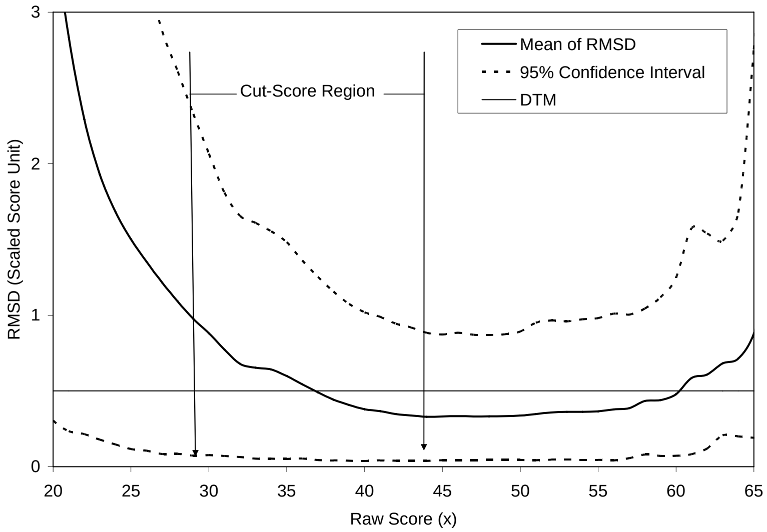

***Figure 11.* The 95% confidence interval (CI) band for the scaled-score unit RMSD: MC-only anchor condition.**

Conclusion

The present study investigated the appropriateness of anchor composition in mixed-format tests using subpopulation invariance indices. Gender subpopulations were formed because many studies reveal gender-by-item format interactions, which can affect linking on mixed-format tests. The two subpopulation functions derived using male and female examinees were compared to the total group function. This study examined the conditional RMSD and overall REMSD measures to assess score equatability, along with subpopulation-specific RESD measures in the two anchor conditions. Raw-to-scaled score conversions were obtained using the chained equipercentile method.

The scaled-score unit REMSD and RESD values were larger than the DTM under the mixed-anchor condition, whereas they were smaller than the DTM under the MC-only anchor condition. Initially, this finding might lead one to the conclusion that the MC-only anchor is superior to the mixed-format anchor in maintaining consistency of the linking relationship across subpopulations. The *ew*REMSD value, however, was twice as large for the MC-only anchor case (0.52) as for the mixed-anchor case (0.26). With the mixed anchor, the conditional RMSD values were smaller than the DTM for the scores (including all the cut scores) where many examinees were located. With the MC-only anchor, however, RMSD values were larger than the DTM for the cut-score region but were smaller than the DTM where many examinees were located. The linking transformation from new to old forms was reasonably consistent across males and females under the mixed-anchor condition. With the MC-only anchor, however, the linking function showed a subpopulation dependence that was large enough to merit attention. Taken altogether, the evidence suggests that the mixed-anchor is more suitable for linking the tests used in this study.[5]

In most score equity assessment analyses, evidence of population dependence might suggest the need to re-evaluate test assembly specifications or linking methods. The results of this study, however, suggest a different remedy. Comparison of males and females in general revealed some differences in the linking functions across subpopulations. Differential subpopulation performance on different item formats had an impact on the linking function, although the effect was not great. Regression analyses, coupled with linking results, indicated that the relationship between MC and CR scores differed by gender; therefore, failure to include CR items in the anchor might bias the resulting linking functions. Here, the indicated problem is with different performance levels on CR and MC items. This tendency would increase if the proportion of males to females is heavily unbalanced across the new and old linking samples. It might be useful to carry out some statistical checks to discover which anchor items function differently for male and female subpopulations after adjusting for subpopulation members' differences in ability.

This study has some practical implications, but it also has limitations. First, the sample sizes for the subpopulations were somewhat unbalanced, but that is the nature of the population for the test equated here. Although the bootstrap procedure was used in this application to compute the standard error of RMSD values to determine whether the differences obtained with the original data were due to random error, implementing this procedure as a standard operation may not be practical for testing programs with strict deadlines.

Second, around 500 papers from the old sample were rescored via a trend scoring procedure by new and old sets of raters in the operational scoring context. We used those papers to assess CR scoring consistency over time but not for linking, because a sample of 500 is not large enough to ensure stability of subpopulation invariance indices. Although we found no statistical difference between the two groups of raters, the use of common CR items scored by different raters across new form and old form groups may yield bias (see Kim et al., in press).

Lastly, the correlation between the CR and MC components of the test was not substantial due to the evident multidimensionality of this mixed format test, and the CR portion was twice as large as the MC portion due to the section weight. Under this circumstance, the MC-only anchor would probably not be an adequate choice for satisfactory linkings of the mixed-format test. Thus, this research does not provide an adequate test of the efficacy of MC-only anchors for more unidimensional tests (e.g., tests of mathematics). It does reflect the results, however, that can be expected for tests in which the construct tested by MC versus CR items differs substantially. Such tests are fairly common (e.g., history, English). Moreover, it does reflect a reality of many testing programs of the over-weighting of CR sections as opposed to MC sections. Anecdotal evidence suggests that such weights reflect (a) the greater time spent by examinees on the CR sections, and (b) educators' beliefs in the greater authenticity of CR items as compared with MC items.

Score equity requires that equated scores have the same meaning no matter when or to whom the test was administered. Lack of subpopulation invariance in a linking function indicates that the variation in difficulty of the new and old forms, as assessed through the NEAT design, is not consistent across those subpopulations. The use of subpopulation invariance indices could serve as a quality check to determine which anchor composition would be better to achieve score equity for the mixed-format test. Although some item format effects were evident in this study, the linking relationship was not influenced seriously by the subpopulations used in deriving the linking function, regardless of anchor composition. The RMSD and resampling results, however, indicate that the mixed-format anchor is a better choice than the MC-only anchor for achieving male and female subpopulation invariance. With the MC-only anchor, subpopulation invariance did not hold, particularly below the 10th percentile (raw score of 34), where several cut scores were located. Using real data, this study demonstrates how the subpopulation invariance application can be implemented to enhance the quality of score equating.

References

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

Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995, April). *A comparison of the results from two equatings for performance-based student assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, *16*, 97-96.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*(1), 77-92.

Cohen, A. S., & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22,* 116-130.

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.

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement,* 3*5*(2), 137-154.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education, 14,* 31-57.

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.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381.

Kim, S., Walker M. E., & McHale, F. (in press). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement*.

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

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal Canada.

Livingston, S. A., & Rupp, S. L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning teachers* (ETS Research Rep. No. RR-04-48). Princeton, NJ: ETS.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1992). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement examinations* (College Board Research Rep. No. 92-7; ETS Research Rep. No. RR-93-05). New York: College Entrance Examination Board.

Muraki, E., Hombo, C. M. , & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*(4), 325-337.

Petersen, N. S., & Livingston, S. A. (1982). *English composition test with essay: A descriptive study of the relationship between essay and objective scores by ethnic group and sex* (ETS Statistical Rep. No. SR-82-96). Princeton, NJ: ETS.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36*, 336-346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329-346.

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.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Notes

[1] As a standard operation, trend scoring was implemented to determine whether raters' scoring standards for the common CR items were consistent across the two administrations. Trend scoring is rescoring the same examinee papers across scoring sessions, thus holding group ability constant so that any scoring shift can be detected (Kim, Walker, & McHale, in press; Tate, 1999, 2000). About 500 papers selected randomly from the old-form group were rescored by the same raters who scored responses for the new-form group. Thus, these trend papers had two sets of scores, associated both with the old and the new sets of raters. The statistical analysis for the trend scoring data revealed no substantial scoring difference between the two sets of raters. The scoring standards for CR items were well maintained across the two administrations; thus, the 6 common CR items could be used with confidence in the anchor.

[2] To use the chained equipercentile equating method, 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. Preserving only univariate moments and no bivariate moment in this situation results in a slightly better fit of the marginal distributions than when the first bivariate moment is also preserved. Such a strategy is possible here because chained equating operates only on the margins. In any event, differences in equating results with and without preserving the bivariate moment are negligible.

[3] Equal weight is the inverse of the total number of scores; thus, in this case the weight is 0.06.

[4] Not all stakeholders use the same cut scores for the test.

[5] We also calculated all deviance measure at the raw-score level. Although the patterns were similar, the magnitude of those measures was generally lower at the raw-score than at the scaled-score level. We can make these results available on request.