Listening. Learning. Leading.®

Research Report
ETS RR–11-44

Does Linking Mixed-Format Tests Using a Multiple-Choice Anchor Produce Comparable Results for Male and Female Subgroups?

Sooyeon Kim

Michael E. Walker

December 2011

Does Linking Mixed-Format Tests Using a Multiple-Choice Anchor Produce Comparable Results for Male and Female Subgroups?

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

December 2011

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

Technical Review Editor: James Carlson

Technical Reviewers: Mary Grant and Gautam Puhan

.

Abstract

This study examines the use of subpopulation invariance indices to evaluate the appropriateness of using a multiple-choice (MC) item anchor in mixed-format tests, which include both MC and constructed-response (CR) items. Linking functions were derived in the nonequivalent groups with anchor test (NEAT) design using an MC-only anchor set for 4 mixed-format licensure tests. For each of those licensure tests, the linking functions were also derived separately for males and females, and those subpopulation functions were compared to the total group function. The mathematics, social studies, and science tests each produced acceptable differences between each of the subpopulation functions and the total group function within the cut-score region, leading to consistent pass/fail designations for the examinees. The English test, which had a low correlation between MC and CR components (indicative of multidimensionality), produced the largest differences, casting doubt on the effectiveness of the MC-only anchor.

Key words: population invariance, mixed-format test, multiple-choice anchor, linking

Table of Contents

List of Tables

List of Figures

Page

Mixed-Format Tests

Many large-scale testing programs include both constructed-response (CR) and multiple-choice (MC) items in their assessments. As with other standardized tests, these mixed-format tests must be equated to ensure equivalence of scores across test forms. Equating most often 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.

The 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. For example, in many cases, CR items are not reused across different test forms because it is easy to memorize and disclose them (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 (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). 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, 2010; 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.

In some cases, the MC-only anchor may in fact represent well the content of a mixed-format test. Evidence for such cases includes a high correlation between the MC and CR portions of the test, such that the disattenuated correlation (i.e., estimated true-score correlation) approaches unity. In other words, the MC-only anchor should be sufficient when the mixed-format test is essentially unidimensional. Walker and Kim (2009) suggested that the MC-only anchor may be effective even when the MC and CR portions of the test appear to fall along somewhat different dimensions: that is, when the disattenuated MC-CR correlation is less than 1. They argued that as long as the function relating CR to MC scores was the same in all examinee groups, then an MC-only anchor would suffice to equate the tests. One rationale for this claim is that if the functional relationship between MC and CR scores remains constant across groups, then the old and new forms should differ in difficulty in the same way across groups for both the

MC and the CR portions of the test. In such a case, an MC-only anchor that adjusts for differences in group ability on the MC portion would adjust equally well for the CR portion. As long as the functional relationship between MC and CR scores displays population invariance, the linking relationship determined in the combined examinee group should hold for all examinee subgroups. The authors called this condition constant dimensionality; they argued that constant dimensionality and not unidimensionality was the necessary condition for successful linking of a mixed format test with an MC-only anchor.

The Use of Population Invariance Indices

In this study, we used subpopulation invariance indices as a way to assess the effectiveness of an MC-only anchor for linking four mixed-format tests. Theoretically, the population invariance requirement means that the equating function must operate independently of subpopulations of examinees from whom the data were 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 old (reference) test forms have not been equated and the interchangeability of the linked scores is questionable (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 salient 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 linked must remain constant across subpopulations. Any inconsistencies in the relationship between anchor and total scores across subpopulations would manifest themselves as subpopulation dependent linking functions.

Recently, Kim and Walker (2009, in press) used subpopulation invariance indices to examine the appropriateness of anchor composition in a large-scale mixed-format licensure test. The authors compared two types of anchor sets: (a) MC only and (b) a mix of MC and CR to see which anchor composition was more likely to result in invariant linking functions. They also examined gender

group by item-format interactions and their direct effects on the linking process. 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 subpopulation dependence, particularly for the cut-score region. Based upon the results, the authors concluded that the mixed-format anchor was a better choice than the MC-only anchor for linking the licensure test, in which the correlation between the CR and MC components was low (r = .43 to .44). The impact of the group by item-format interaction on the linking function was minimal.

Purpose

In this study, we examined the appropriateness of using an MC anchor set in linking mixed-format tests using subpopulation invariance indices. For this purpose, we selected four large-scale licensure tests measuring different academic subjects (English language arts, mathematics, social studies, and science), which tend to show different levels of correlations between MC and CR components due to the nature of the constructs being measured. In practice, MC-only anchors have been used for those tests because common CR items were unavailable. The contribution of the CR component to the composite score was relatively small (25%), and relatively few CR items per form were included; thus, CR items were not reused across forms.

We tested for invariance using gender subpopulations because many studies reveal gender by item format effects (Livingston & Rupp, 2004; Mazzeo, Schmitt, & Bleistein, 1992; Petersen & Livingston, 1982; Willingham & Cole, 1997), which could potentially affect linking on mixed-format tests. Accordingly, gender subpopulations could play a particularly important role in the linking process for mixed-format tests.

Methods

Data

This study involved the linking of four pairs of test forms through common items. Data sets from two national administrations for each of four large-scale licensure tests were used. The data were collected using a NEAT design. Table 1 describes the structure and scoring of each of the four operational licensure tests: the number of items in the test, the number of anchor items, the proportion of MC and CR components, and the scoring scale and weighting factor for the CR components. For each test, examinees could earn three times as many raw-score points on the MC section as on the CR section. The proportion of MC anchor items (33%) was constant across the

four tests. As shown in Table 2, the number of total examinees ranged approximately from 900 to 1,900 per form of each test, and more than 65% of the examinees for each test were females.

Table 1

General Information for the Four Tests Used in This Study

Test	English language arts	Mathematics	Social Studies	Science
MC items, points per item, scoring weight	90 × 1 × 1	45 × 1 × 1	90 × 1 × 1	90 × 1 × 1
MC points (percent contribution to total score)	90 (75%)	45 (75%)	90 (75%)	90 (75%)
CR items, points per item, scoring weight	2 × 6 × 2.5	3 × 6 × .8333	3 × 6 × 1.6667	3 × 6 × 1.6667
CR points (percent contribution to total score)	30 (25%)	15 (25%)	30 (25%)	30 (25%)

Note. Two problematic MC items were not scored in the new form of the English and social studies tests and in the old form of the science test. Therefore, the actual possible composite score point was 118 rather than 120 for those test forms. CR = constructed response, MC = multiple-choice.

Table 2

Total Number of Examinees, Proportion of Each Gender, and Standardized Mean Difference Between Male and Female Subgroups (Male Minus Female) for the Four Tests

Test score	English	Mathematics	Social studies	Science
New form				
Total *N*	1223	1875	881	1224
Male (%)	16.5%	26.1%	34.2%	30.3%
Female (%)	83.5%	73.9%	65.8%	69.7%
Composite (M-F)	-0.15	0.07	0.55	0.55
MC anchor (M-F)	0.04	0.09	0.59	0.46
Old form				
Total *N*	1445	1054	1197	911
Male (%)	15%	28.9%	32.2%	28.1%
Female (%)	85%	71.1%	67.8%	71.9%
Composite (M-F)	-0.09	0.04	0.57	0.35
MC anchor (M-F)	-0.06	0.12	0.64	0.32

Note. A negative sign indicates that the female subgroup performed better than the male subgroup on that measure. F = female, M = male, MC = multiple-choice.

Procedure

In the NEAT design, the linking function derived using each subgroup was compared to the linking function derived using all examinees to determine whether the total group linking function would yield scores comparable to the subpopulation linking functions. The study included two steps.

Step 1. **Obtain total group and subpopulation linking functions using an MC-only 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-linked raw-score conversions were obtained using the chained equipercentile linking method.[1] The linking process determines, for each possible raw score on the new form, the corresponding raw score on the old form.

Step 2. Compare total group and subpopulation linking functions. The linking function derived using each subpopulation was compared to the total group linking function. The differences were quantified across all subgroups using three deviance measures: the root mean square difference (RMSD) at each score level on the new form (von Davier, Holland, & Thayer, 2004a); the root expected mean square difference (REMSD; Holland, 2003), which is a weighted average across score points of the RMSD; and the equally weighted root expected mean square difference (*ew*REMSD), which is an unweighted average of the RMSD (Kolen & Brennan, 2004, p. 443).

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. 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 a raw-score point in the raw-to-raw score transformations, was used. Half a point was used here because we could reasonably expect any differences of less than half a point to round to the same integer score value. The *ew*REMSD, which gives equal weight to all score points, was also calculated. We paid particular attention to the raw cut-score region[2] to examine the impact of subpopulation influence on the examinees' pass/fail designations.

As shown in Table 2, the subpopulation sizes were seriously unbalanced (more than 65% female). Because the female subpopulation heavily influenced the RMSD and REMSD measures, we separately quantified the difference between subpopulation linking functions and the total group linking function to assess more clearly the impact of reporting scores for that subpopulation based on the total group linking transformation. To do so, we used the root

expected square difference (RESD), a single value weighted index of the difference between the conversions computed on the subpopulation and the total population. Thus, each subgroup would have a single summary RESD value. All deviance measures were computed based on unrounded scores. In general, a negligible difference indicates that the linking relationship is not influenced significantly by the subpopulations used in deriving that function.

The formulas for calculating RMSD, REMSD, *ew*REMSD, and RESD are as follows:

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

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

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

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

where x represents each raw-score point, $e_{yij}(x)$ indicates the equipercentile linking function in the *j*th subpopulation for score point *i*, $e_{yi}(x)$ 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. The proportional (i.e., unequal) weight derived from the actual relative size of each subpopulation was imposed on each subpopulation when calculating RMSD, REMSD, and *ew*REMSD.[3]

Steps 1 and 2 were repeated in the same manner for each of the four subject tests. Furthermore, Steps 1 and 2 were repeated using scores of the MC component (excluding the CR component) to determine whether the MC-only anchor could successfully place the new and old MC scores on the same scale, achieving adequate degrees of population invariance. The effectiveness of an MC-only anchor for the mixed-format test would be questionable if the MC anchor set yielded invariant linking functions for the MC component of the test but not for the composite of MC and CR components. On the other hand, if the MC component failed to demonstrate population invariance when linked with a particular MC anchor, this finding would call into question the quality of that particular anchor and not necessarily the viability an MC-only

anchor. Comparison of the MC component versus the composite score linking results improves detection of the source of composite score equating failures if they arise.

Results

Preliminary Analysis

Table 2 presents the standardized mean differences between male and female subgroups on the composite and MC anchor scores for each subject test. Figure 1 displays the same type of difference on the total MC and CR scores for each test. For the English test, females were higher in ability than were males as measured by either MC or CR, or both. The gender difference was much larger on the CR component than on the MC component of the test in the new-form group. For the social studies and science tests, males performed much better than did females on all measures, but the gender difference was much larger on the MC component than on the CR component. The gender difference was minimal for the math test. Gender differences were fairly constant across the new- and old-form groups except for the English test.

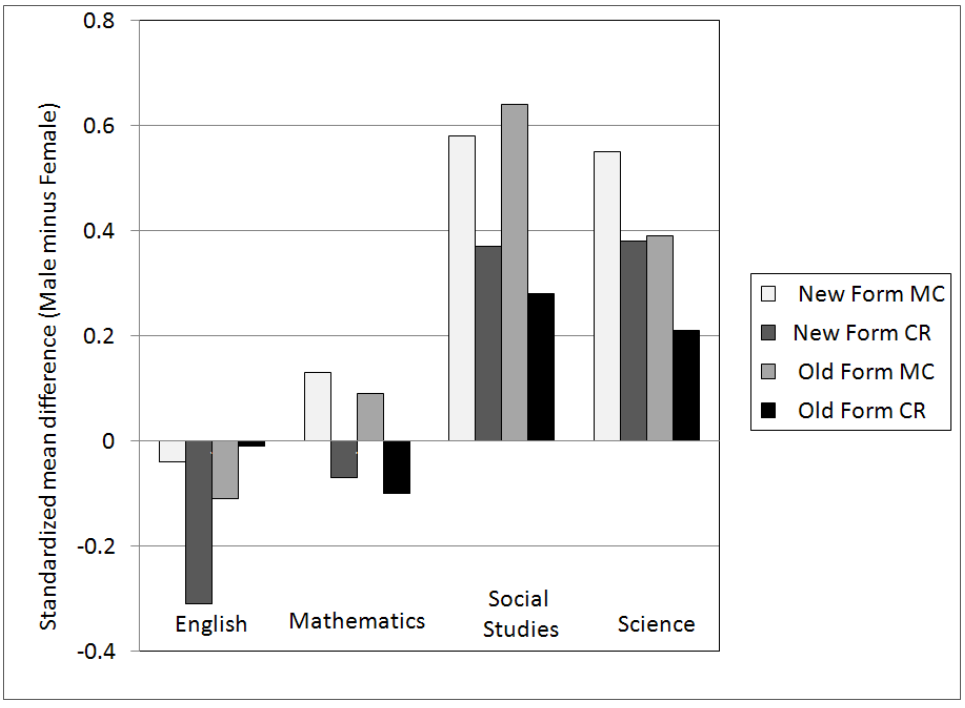

Figure 1. **Standardized mean difference between male and female subgroups on the multiple-choice (MC) and constructed-response (CR) scores.**

Table 3 presents correlations of MC scores with CR scores for the four tests in the new- and old-form groups, along with disattenuated (i.e., estimated true-score) correlations based on the computed total test reliability coefficients. For the English test, the correlation was lowest and the relationship between MC and CR components was least consistent across the new- and old-form groups. Such low disattenuated correlations indicate that MC and CR components measure somewhat different constructs. The multidimensionality of the test was evident. The math and social studies tests showed fairly consistent relationships between the two components, although their disattenuated correlations were less than unity. Only for the science test did the disattenuated correlations approach unity, providing some evidence for the unidimensionality of this test.

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 and Figure 1 show, the gender difference was somewhat inconsistent as a function of item format for the tests. To examine the interaction effect in more detail, moderated regression analyses were performed predicting CR scores from MC scores, gender (male vs. female), and the MC by gender interaction for each of the eight conditions (two administrations × four tests). Seven of eight conditions displayed statistically nonsignificant gender by item format interaction effects. The old form for English showed a small but statistically significant MC by gender interaction. The interaction did not appear sizable enough to have an appreciable impact on the linking results. These findings are consistent with those of Kim and Walker (2009, in press).

Table 3

Correlations Between Multiple-Choice (MC) and Constructed-Response (CR) Scores for the Four Tests in the New- and Old-Form Groups

Test form	Group	Correlations between MC and CR			
		English	Mathematics	Social studies	Science
New form	Male	.33 (.54)	.65 (.87)	.63 (.88)	.67 (.97)
(*X*)	Female	.40 (.67)	.65 (.87)	.63 (.92)	.65 (.98)
Old form	Male	.49 (.73)	.62 (.83)	.60 (.84)	.65 (.98)
(*Y*)	Female	.47 (.68)	.63 (.87)	.64 (.88)	.72 (1.00)

Note. The correlations presented in parentheses indicate disattenuated (i.e., estimated true-score) correlations of MC scores with CR scores based on classical true-score theory.

Subpopulation Linking Analysis

To determine if the manifest differences in linking functions across subpopulations corresponded to the constructs being measured, each new form was linked to the corresponding old form for the total group and for the two subpopulations. Deviance measures among the resulting raw-to-raw score conversions were compared to the DTM to assess the extent of the differences among the conversions. Table 4 displays a summary of population invariance indices for the four mixed-format tests for the MC and composite scores. Figures 2 to 5 present the RMSD results derived from linking MC scores through MC-only anchor scores, and Figures 6 to 9 present the same information derived from linking composite scores through MC-only anchor scores.

Table 4

Summary of Raw-Score Population Invariance Indices for the Four Mixed-Format Tests

Indices	English	Mathematics	Social studies	Science
Total MC score				
REMSD	.128	.268	.248	.274
RESD: Male	.285	.455	.393	.466
RESD: Female	.065	.142	.123	.129
Composite (MC plus CR scores)				
REMSD	.715	.365	.385	.523
RESD: Male	1.617	.590	.412	.755
RESD: Female	.342	.228	.371	.389
*ew*REMSD	.810	.243	.380	.362

Note. CR = constructed response, MC = multiple-choice, REMSD root expected mean square difference, RESD = root expected square difference, *ew*REMSD = equally weighted root expected mean square difference.

The data presented in Table 4 indicate that, when MC total scores were linked using MC anchors, the summary REMSD values were much smaller than the DTM. The differences between the subpopulation and total population conversions appear to be negligible for all four tests. Figures 2 to 5 present the conditional raw-score unit RMSD along with the summary REMSD and DTM for all new-form MC raw-score points above the first percentile for the English, math, social studies, and science tests, respectively. For all the tests, the linking functions derived from male and female subpopulations did not differ from the total group linking function across the raw-score points where most examinees were located. In other words,

the derived linking function did not depend upon choice of linking (sub)population. The negligible population dependence of the linking functions provides evidence that the total MC scores were successfully equated via the MC anchor.

Table 4 and Figures 6 to 9 present deviance measures associated with linking of composite scores using MC anchors. Again, the figures show information for all new-form composite score points above the first percentile. As expected, all four tests yielded larger RMSD and REMSD values for linking composite scores than for linking total MC scores. For the math, social studies, and science tests, the difference between the total group function and each subpopulation function was outside the DTM range for the low and high ends of the score range, where data points were sparse. For the cut-score region, however, the linking functions derived from male and female subpopulations did not differ appreciably from the total group linking function for any of these three tests, leading to consistent pass/fail designations for the examinees. For math, social studies, and science, the mixed-format test cut scores were successfully preserved using MC-only anchors. Figures 7 and 8 provide evidence that for math and social studies, the majority of scores on the mixed-format tests were successfully equated using MC-only anchors.

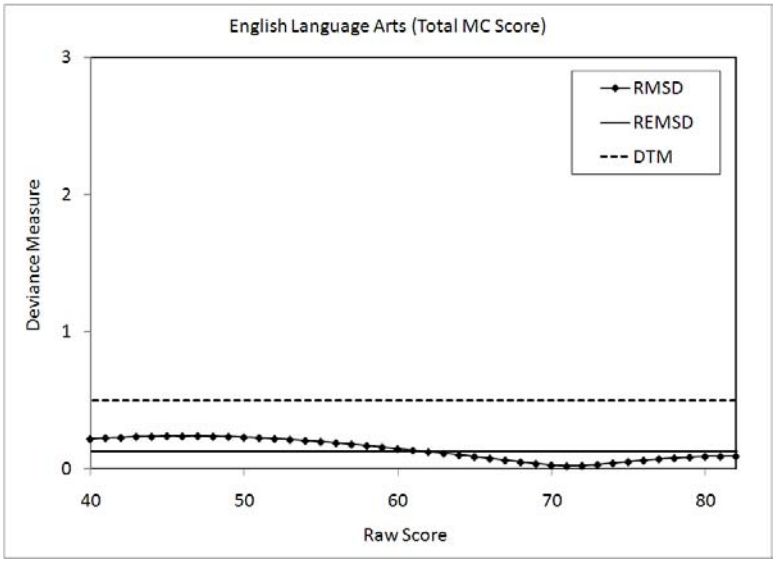

Figure 2. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: English language arts, total multiple-choice (MC) score.**
Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

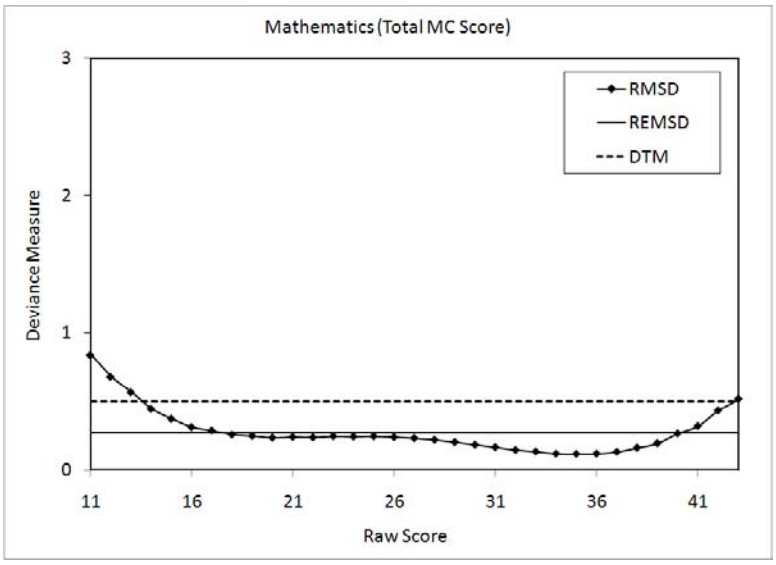

Figure 3. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: mathematics, total multiple-choice (MC) score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

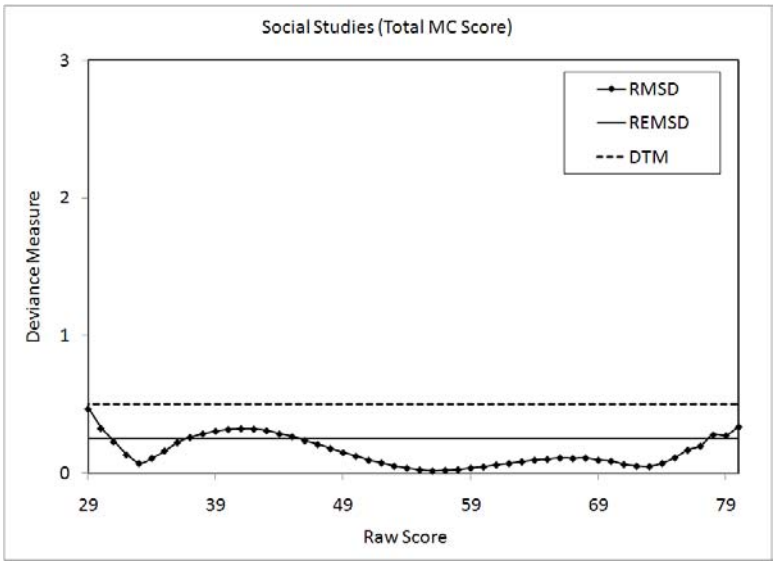

Figure 4. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: social studies, total multiple-choice (MC) score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

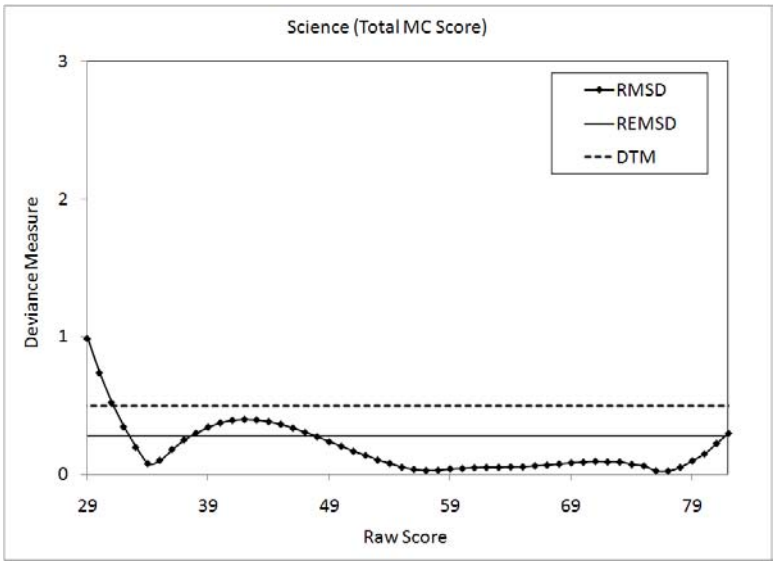

Figure 5. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: science, total multiple-choice (MC) score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

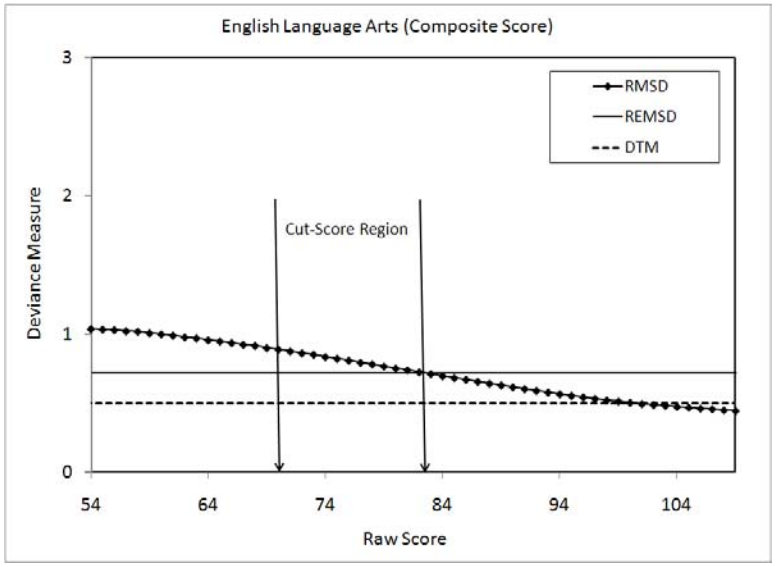

Figure 6. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: English language arts, composite score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

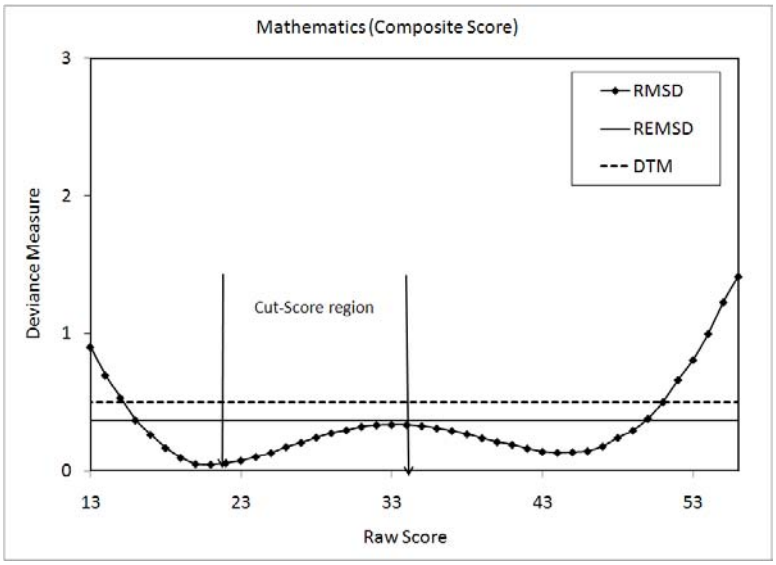

Figure 7. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: mathematics, composite score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

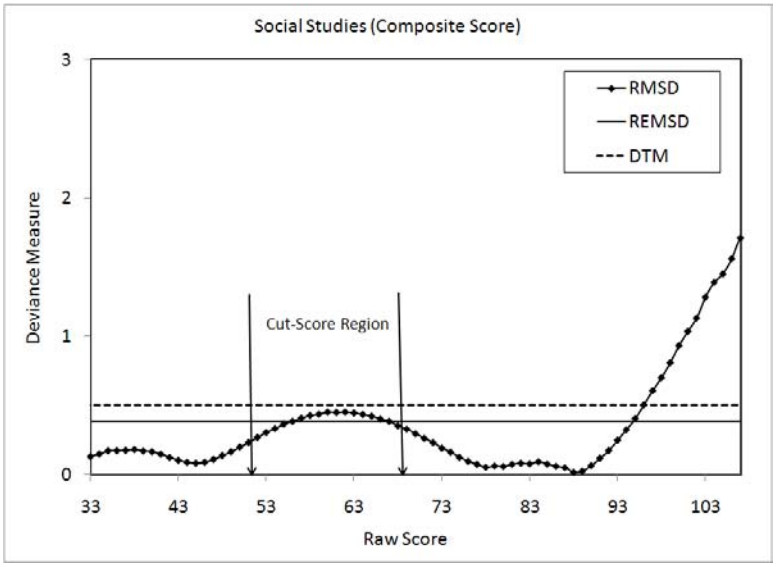

Figure 8. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: social studies, composite score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

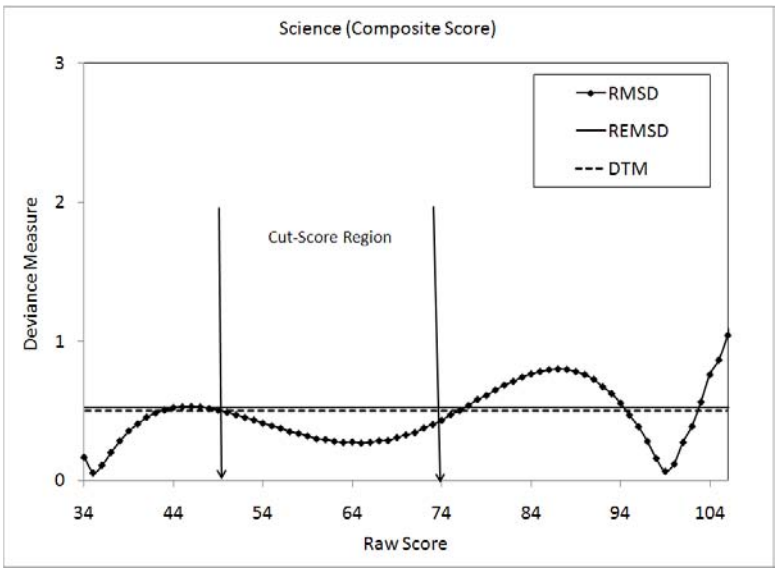

Figure 9. **Raw-score deviance measure curves comparing the subpopulation linking functions with the total group linking function: science, composite score.**

Note. DTM = difference that matters, REMSD = root expected mean square difference, RMSD = root mean square difference.

If for math, social studies, and science, the population dependence of linking functions was not large enough to seriously affect pass/fail designations, the same cannot be said for the English test. The deviance curves in Figure 6 indicate that the linking functions derived from male and female subpopulations differed substantially from the total group linking function across the entire score region. The difference between the total group function and each subpopulation function fell outside the DTM range, and thus both REMSD and *ew*REMSD values were much larger than DTM. The effectiveness of the MC-only anchor is questionable for this test.

Our investigation included four different test titles and two types of total scores (composite and MC total). To facilitate comparisons among them, we have combined the results for each test into a single graph. We constructed the graph by computing the RMSD values at specified percentile points in the new-form score distribution and expressing the RMSD values in standard-deviation units. Note that both the horizontal scale and the vertical scale of the graph are defined in terms of the score distribution. The RMSD values are conditioned on percentile

points of the distribution and expressed in terms of the standard deviation of the distribution, making it meaningful to compare the RMSD values across the different tests.

Figure 10 depicts the RMSD (expressed in SD units) derived from linking total MC scores through the MC anchor score for each of the four tests at the nine percentiles. When the new MC component was linked to the old MC component through the MC anchor set, the English and social studies tests both evidenced invariant linking functions across the score range. The math and science tests revealed a subgroup dependency at the extremes of the score distributions; this dependency, however, may not have practical significance because all tests' cut scores fell between the 10th and 75th percentiles.

Figure 11 presents the same comparison as Figure 10, but with composite scores. The English test produced the largest RMSD for raw scores below the 75th percentile, where most cut scores were located. This finding is not surprising because the English test showed no evidence of unidimensionality. Here, the indicated problem for the English test is with a group of individuals with different performance levels on CR and MC items, potentially indicating lack of constant dimensionality across the old- and new-form groups. For the remaining three tests, reasonable invariance of linking functions across gender subpopulations was attained for most low- and middle-performing examinees but not for the highest performing examinees.

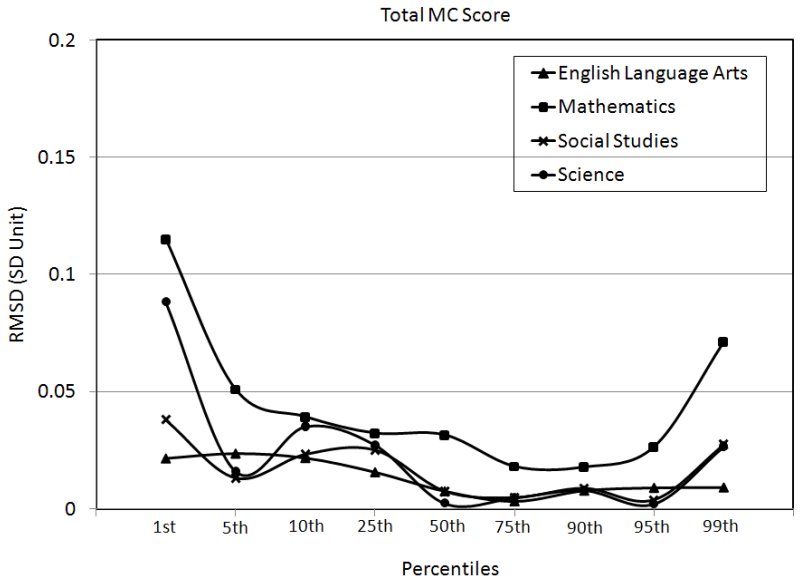

Figure 10. **Root mean square difference (RMSD) SD unit derived using multiple-choice (MC) total scores.**

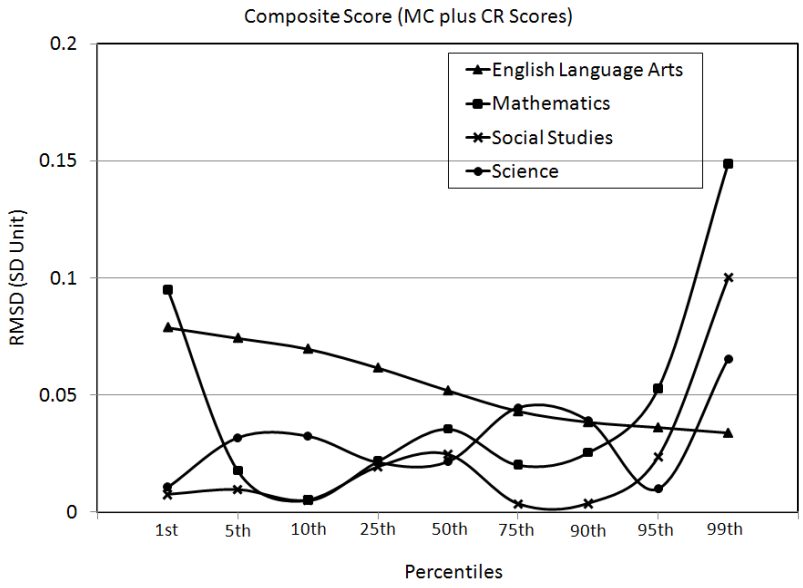

Figure 11. **Root mean square difference (RMSD) SD unit derived using composite scores.**
Note. CR = constructed response, MC = multiple-choice.

Conclusion

We used subpopulation invariance indices to investigate the appropriateness of MC-only anchor composition in mixed-format tests. We selected four academic subject tests to determine if the effectiveness of the MC anchor would depend on the relationship between the MC and CR components. Our initial prediction was that MC and CR items are more likely to measure somewhat different constructs in the English and social studies tests but similar constructs in the mathematics and science tests, due to the nature of the constructs being measured. Gender subpopulations were investigated because many studies reveal gender by item format interactions, which can affect linking on mixed-format tests.

For all four mixed-format tests, MC components were successfully linked through the MC anchor, maintaining the same old- to new-form relationship across gender subpopulations. Although subgroup invariance was somewhat questionable for low-performing examinees, the degree of subgroup dependency was extremely small. Interestingly, among the four tests, the English test achieved the best level of population invariance for linking MC components through MC anchor scores. When composites of MC and CR scores were linked through MC-only anchors, however, the linking functions for the English test were subpopulation dependent. As

the comparisons among the RMSD values for the four tests indicated, the English test was inferior to other mixed-format tests in maintaining reasonable consistency of the linking relationship across subpopulations using the MC-only anchor. The mathematics, social studies, and science tests attained reasonable levels of subpopulation invariance, especially in the cut-score region, which is most important for licensure tests. For these mixed-format tests, evidence suggested that the linking of the cut scores on the new composite to those on the old composite through the MC anchor could reasonably be called equating. That is, the cut scores across forms could be treated as exchangeable without detrimental effects. Exchangeability would not necessarily hold across the entire score scale.

In the context of a NEAT design, we expect linking methods to be successful if the anchor and the composite scores measure the same construct. For mixed-format tests, if the MC and CR portions measure the same construct, in principle we would expect an MC-only anchor (or a CR-only anchor, for that matter) to be sufficient to equate the test forms. As a simple test of the unidimensionality of the mixed-format tests in this study, we computed the disattenuated correlation between the MC and CR sections. This correlation approached unity for the science test, which is an essential characteristic of two measures of the same construct (Brennan, 2007). For this test, then, we would expect an MC-only anchor to suffice for linking. The other tests had disattenuated MC-CR correlations of much less than unity, casting doubt on the unidimensionality of these mixed-format tests.

Von Davier, Holland, and Thayer (2004b, pp. 36–37) listed two assumptions that must be met for chained equipercentile methods to yield equated scores. The first is that the (equipercentile) relationship between the anchor and total scores on the old form is invariant across populations. The second is that the relationship between the anchor and total scores on the new form is invariant across populations. If these two conditions hold, then the chained equipercentile linking should be invariant across populations. These invariance conditions should be met when the anchor test is construct representative of the total test. In the case of an MC-only anchor and a mixed-format test, the anchor can be construct representative of the total test only to the extent that the MC and CR portions measure the same thing (i.e., the test must be unidimensional).

Walker and Kim (2009) considered that unidimensionality of MC and CR portions may not be necessary to achieve population invariant linking with an MC-only anchor. Suppose that

the MC and CR portions measured somewhat different constructs, such that their disattenuated correlation is less than 1.0. Suppose, though, that the functional (e.g., equipercentile) relationship of the CR to the MC scores remained invariant across populations. If the relationship between the MC-only anchor and the MC scores were population invariant, and the relationship of the MC to CR scores were also invariant across populations, then the relationship between the anchor and the MC-CR composite scores would necessarily be invariant as well. The invariance of this anchor-total relationship qualifies chained equipercentile linking as an equating. The invariance of the MC-CR relationship, which the authors called constant dimensionality, can be seen as a necessary condition for chained equipercentile linking of a mixed-format test via an MC-only anchor to qualify as an equating; and unidimensionality can be seen as a special case of this broader condition.

One way to test the assumption of constant relationships across groups is to compare the regressions of CR scores on MC scores across different groups. The moderated regressions did not show large interactions between MC score and gender. However, sizable gender main effects manifested themselves for the old form of English. Smaller gender effects were seen for the science old form and the social studies new form. The relative magnitudes of these differences are roughly reflected in relative sizes of the deviance measures in Figures 6 through 9 and 11. At this point these findings provide perhaps only indirect support for the constant dimensionality assumption. Still, they suggest a simple tool for investigating the effectiveness of MC-only anchors for different mixed-format tests.

In other uses of subpopulation invariance analyses, population dependence might suggest the need to re-evaluate test assembly specifications or possibly the linking method. Here, the indicated problem is neither with the test specifications nor with the linking method, but with a group of individuals with different relative performance levels on CR and MC items. The problem might be exacerbated if the proportion of males to females were heavily unbalanced across the new and old linking samples; this was the case in the present study. 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. The subgroup membership was heavily unbalanced for all tests, with 65% or more of the examinees being female. Ironically, this imbalance may eliminate some subgroup dependency concerns because in

practice the impact of subgroup dependency would influence only a small number of examinees. The generalizability of the current findings may be limited. We used data sets collected in actual operational settings. Accordingly, many psychometric factors (e.g., the size of the correlation between MC and CR, score differences between males and females on MC and CR items, the proportion of MC and CR components) were not controlled. A simulation study would be appropriate to allow manipulation of various psychometric conditions to clarify their effects.

To make a claim of group invariance, the resulting equated scores should have the same meaning no matter when or to whom a test is administered. This condition would be more likely to hold when the composition of the anchor matches those of the test forms to be equated. For mixed-format tests, excluding CR items from the anchor set is quite common because of a reluctance to readminister such memorable items. Subpopulation invariance may not hold if the anchor set is unrepresentative in terms of content coverage and psychometric properties. The use of subpopulation invariance indices could serve as a quality check to determine whether an MC-only anchor set would be sufficient to achieve an invariant equating function for a mixed-format test. If linkings across subpopulations lead to the same relationship between old and new forms, use of the MC-only anchor would be supported. Using real data, this study demonstrated how subpopulation invariance indices could be used to enhance the quality of linking functions. The study suggested a moderated regression procedure that might yield similar information without the need to perform multiple linkings; as such, this simpler procedure merits further attention.

References

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

Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995). *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.

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

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An explanatory study in an advanced placement history examination. *Journal of Educational Measurement, 31*, 275–293.

Brennan, R. L. (2007). Tests in transition: Discussion and synthesis. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). New York, NY: Springer-Verlag.

Dorans, N. J., & Feigenbaum, M .D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, 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, pp. 93-124). Princeton, NJ: ETS.

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

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, 35*, 137–154.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195–208.

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 Report 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*, 357–381.

Kim, S., & Walker, M. E. (2009). *Evaluating subpopulation invariance of linking functions to determine the anchor composition for a mixed-format test* (ETS Research Report No. RR-09-36). Princeton, NJ: ETS.

Kim, S., & Walker, M. E. (in press). Determining the anchor composition for a mixed-format test: Evaluation of subpopulation invariance of linking functions. *Applied Measurement in Education*.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement, 47,* 36–53.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: 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 Report 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 Report No. 92-7; ETS Research Report No. RR-93-05). New York, NY: College Entrance Examination Board.

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

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

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

Walker, M. E., & Kim, S. (2009, April). *Linking mixed-format test using multiple choice anchors.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Notes

[1] The data were pre-smoothed 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 moments 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.

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

[3] We also calculated all deviance measures using the equal weight (i.e., 0.5 for each subpopulation). In general, the equal weight condition produced slightly greater RMSD than did the proportional weight condition, but the patterns were very similar and the difference between the two weights was almost negligible. We can make these results available on request.