Measuring the Power of Learning.™

Research Memorandum
ETS RM–19-08

Equating a Mixed-Format Test With and Without Constructed-Response Anchor Items

Chi-Wen Liao

Wei Wang

Yi Cao

August 2019

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Equating a Mixed-Format Test With and Without Constructed-Reponse Anchor Items

Chi-Wen Liao, Wei Wang, and Yi Cao
Educational Testing Service, Princeton, New Jersey

August 2019

Corresponding author: C.-W. Liao, E-mail: CLiao@ets.org

Suggested citation: Liao, C.-W., Wang, W., & Cao, Y. (2019). *Equating a mixed-format test with and without constructed-response anchor items* (Research Memorandum No. RM-19-08). Princeton, NJ: Educational Testing Service.

Action Editor: Shelby Haberman

Reviewers: Sooyeon Kim and Samuel Livingston

Abstract

In this study, we investigated the effectiveness of different anchor designs for equating a mixed-format test that contains multiple-choice (MC) and constructed-response (CR) items. In the test, the MC score makes up most of the total test score. The criterion equating employed an MC/CR anchor with rescoring of CR anchor responses of the reference-form test takers. Using the MC/CR anchor without rescoring produced large equating error. Equating through an MC-only anchor produced smaller conditional equating biases than the MC/CR anchor without rescoring of the CR responses in the part of the score range where the majority of test takers scored, yet it still yielded an equating error greater than half a raw score point. The equating that employed only 1 scorer in the CR anchor rescoring produced similar results to the equating that employed 2 scorers; however, we observed a trend that the new-form conversion tended to be too high for low-scoring test takers and too low for high-scoring test takers.

Key words: equating, constructed-response items, scoring shift, mixed-format test, rescoring process, trend scoring, equating bias

A mixed-format test contains both multiple-choice (MC) and constructed-response (CR) items. The inclusion of CR items in the test is intended to enhance the authentic representation of real-world problems. The response format of CR item types can be quite versatile, ranging from short written responses to long essays. Although the inclusion of CR items has been popular, the task of equating a mixed-format test across administrations is much more complicated than equating a test that contains only MC items.

When researchers equate a mixed-format test using the common-item nonequivalent groups design, it would be easier to achieve an optimal result if they could construct an anchor test that represents the content and statistical properties of the total test. However, depending on the number of CR items and the scoring weight the items carry, an ideal anchor may be difficult or impossible to achieve. A mixed-format test typically contains only a few CR items. On such a test, it is sometimes difficult to make the content and statistical properties of the anchor representative of the total test. Also, some types of CR items—for example, a simple essay prompt—tend to be remembered by test takers. If the CR item is repeated in a new form, test takers may have advance knowledge of it, and using it in the anchor will introduce a bias into the equating. Furthermore, human scoring of CR items is prone to change over time (Bock, 1995; Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). Even if an ideal anchor test, including CR items, could be successfully constructed, unintended differences in the scoring standards of the scorer groups at different administrations complicate the issue of how to accurately equate a mixed-format test. When scoring standards change, the estimate of group differences on the basis of the CR common items is contaminated because the group-to-group difference statistic is no longer a pure reflection of differences in the test takers' ability.

Tate (1999) cautioned that inconsistent scoring standards across administrations would cause inaccurate equating results and proposed a special linking study method to control the scoring shift across administrations. He conducted the equating for an all-CR test in the framework of item response theory using a partial credit model, and this special linking study called for scorers in the second administration to rescore the CR responses of a random sample of test takers from the first administration. This way, the CR responses of this sample could be directly compared with those of the new-form test takers. Tate (2000) further conducted a simulation study to examine the effectiveness of the proposed approach for mixed-format tests and found that using an MC-only anchor in equating a mixed-format test that had an MC/CR

correlation of .60 led to a large bias in the linking coefficient estimates. When CR makes up only a small portion of the total score, it is reasonable to expect that the impact of a CR scoring shift on equating, if present, will be small. The question is, how small does the CR portion need to be? When the MC and CR items measure closely similar skills, it is reasonable to use an MC-only anchor to equate the test. The problem is, how high does the correlation between MC and CR portions need to be?

Lee, He, Hagge, Wang, and Kolen (2012) fixed their MC and CR proportion to be 50% of the total score in their test and investigated the effectiveness of the MC-only anchor in various degrees of MC/CR correlation, ranging from .50 to 1.00. To achieve an overall equating error less than .50 raw score points, they found that the MC and CR portions need to be correlated at least .70 when the means of the anchor scores between the new-form and reference-form groups differed by .10 standard deviations. Using similar conditions, Wang (2013) found that the minimum MC/CR correlation could be as low as .50. Walker and Kim (2010) used a test in which the MC made up only one third of the total score. They did not report the correlation of the MC and CR scores, but the MC anchor and total score correlation was approximately .55. The results showed that the MC-only anchor design yielded a very large equating bias. Kim, Walker, and McHale (2010) also used a test in which the MC made up one third of the total score and did not find the MC-only anchor to be effective. In their study, using the MC-only anchor produced about three times the root mean square error (RMSD) as compared to the MC/CR anchor with rescoring of CR items. Because the CR portion of the test used in their study made up two thirds of the total score, we wondered whether their results could be generalized to tests in which CR makes up only one-third or less of the total score. Would the conclusion change with a different test in which the MC items make up a much larger portion of the total test score and the MC/CR correlation is moderate? These inquiries motivated the current study. The first purpose of the current study was to evaluate the effectiveness of two anchor designs when equating a mixed-format test in which the MC portion makes up approximately two-thirds of the total score and the MC/CR true correlation is moderate.

Rescoring CR items is costly and time consuming and may not be practically feasible for many testing programs. We posit that both the additional cost and time of rescoring the CR items would be greatly reduced if the rescoring were done with only a single scoring of each response instead of the two scorings used in the operational scoring. The second purpose of the current

study was to examine the effectiveness of using one scorer instead of two to rescore CR responses if a rescoring process is necessary. In operational settings, CR items are typically scored by two scorers to reduce human error and increase test-score reliability. If two scorers had been scoring very similarly, single scoring would reduce the reliability of the CR scores, but the reduction should be small compared with double scoring. Kim and Moses (2013) assessed the psychometric conditions under which single scoring for CR items was as effective as double scoring in the context of licensure testing. They studied five mixed-format tests that varied in the MC/CR proportions and interrater correlations. They found that test takers' pass/fail percentage agreement decreased as the CR portion of the total score increased and interrater correlations decreased. Typically, the more the total score is made up of the CR portion, the less reliable the total score will become. Switching to a single scoring model would possibly worsen the problem. Kim and Moses found that single scoring reduced the composite score (i.e., the total score of MC and CR) reliability by only .02 for tests in which the contribution of the CR portion was as small as 25% to 33% of the total score and by .04 or more for tests in which the contribution of the CR portion was 50% or more. They found support for the transition to single scoring from double scoring for tests with a CR component as low as one third of the total score and about a .50 interrater correlation among the CR items in the test. In our study, the CR items contributed one third of the total score points. The operational scoring of CR items occurs in an online-distributed scoring setting (i.e., scorers work remotely from their own computer terminals, which are not located a single scoring center), and scorers are randomly assigned to score CR responses. Both scores of a response are made by scorers from the same scorer pool. There seems to be no reason to expect systematic differences between the first and second scores of the same responses. The second purpose of the study was thus to investigate the effect of having only one scorer rescore CR responses in the reference-form rescored sample, while the operational CR scoring of the new-form responses continued to be done by two scorers. The rescored sample is a random sample of the total reference-form group.

In the first part of the study, we sought to evaluate the effectiveness of two anchor designs for equating a mixed-format test. The two anchor designs were an MC-only anchor and an MC/CR anchor without rescoring of the reference-form test takers' responses to the CR anchor items. The assumption in the MC-only anchor design is that the groups taking the new form and reference form will differ just as much in the CR skills as in the MC skills. The

assumption in the MC/CR anchor design with no rescoring is that no scoring shift has occurred; the effective CR scoring standards for new-form test takers were the same as for the reference-form test takers. The purpose was to answer the question, do we need to include CR items in the anchor, and if so, do we need to rescore the reference-form test takers' responses to those CR anchor items?

In the second part of the study, we investigated the possibility of equating through an MC/CR anchor with a partial rescoring of the reference-form test takers' responses to the CR anchor items. In this case, the partial rescoring consisted of a single scoring of each response, instead of two independent scorings from two scorers with adjudication of any large differences. The purpose of this analysis was to see how closely the results of the partial rescoring matched the results of the full rescoring, to answer the question, if we do need to rescore CR responses of the reference-form test takers' responses, is a single scoring of each response sufficient?

The motivation for conducting these investigations came from inquiries that we continually encounter from administrators of the testing program. Some questioned the necessity of trend scoring CR items when the CR score makes up only about 30% of the total score. Others asked about ways to reduce the cost of CR rescoring if it is necessary. The test is used to make high-stakes decisions about licensure for entry-level educators. Its purpose is to assess whether entry-level educators have the knowledge, skills, and abilities needed for entry into the profession. The test is available every month and is taken by approximately 6,000 candidates each year. The primary score users are state licensing agencies. Because this is a licensure test, we also evaluated the consistency of the pass–fail decisions from each of the equating methods with the pass–fail decisions from the criterion equating.

Method

Data

One type of testing situation offers an especially good opportunity to study the intended research questions. It is the reuse of an entire form of a mixed-format test, with the rescoring of a sample of responses from the original administration of the test. Using these data, we can equate the scores on the second administration of the form (used as the new form in this study) to the scores on the original administration of this form (used as the reference form). Because the new and reference forms are the same form, the whole MC section, which makes up 70% of the total score for this test, can be used as an anchor in the equating. Two different forms, Forms A and B,

of the test were each readministered, and the resulting data were used for this study. Before using these data as the basis for this study, we computed the means and standard deviations of the item *p*-values (proportion correct) in the new-form and reference-form test-taker groups and the correlation of the new-form and reference-form *p*-values. The means and standard deviations were nearly identical in the two groups, and the correlations were approximately .99. We also made a scatterplot with a data point for each MC item and a data point for each CR item score threshold, plotting the difficulty in the new-form test-taker group against the difficulty in the reference-form test-taker group. All of the items were close to the diagonal line, and no items appeared as outliers in the plots. (See the plots of the *p*-values and summary statistics from the two groups in Figures A1 and A2 and Tables A1 and A2.)

Test

The test contains 80 MC items and 7 CR items. The operational scoring of the CR items occurs in an online-distributed scoring setting. Each CR response is independently scored by two scorers, scoring holistically on a scale of 0 to 3. Adjudication is triggered when the discrepancy between the first and second scoring of a response is greater than 1 point on the 0–3 scoring scale. The final CR score is the sum of the adjudicator's score and the original score that is closer to the adjudicator's score. If the adjudicator's score is halfway between the first and second scores, the final CR score is 2 times the adjudicator's score. The CR items require test takers to demonstrate professional knowledge by providing written, in-depth explanations and solutions for problems presented in the item prompt. Test takers are required to analyze scenarios of problems and sets of documents and data to propose an appropriate course of action and rationales for their proposals. The CR scores are combined with the MC scores to form a raw total score, using weights that makes the MC and CR scores 70% and 30% of the total score. The maximum possible total score is 114 for Form A and 113 for Form B. The disattenuated correlation between the MC and CR scores of this test is typically around .70, indicating that the two scores may be measuring somewhat different aspects of the same construct. The operational CR interrater correlations of this test were typically around .50 and higher for items across forms.

Analysis, Part I

To test the assumption that the groups taking the new and reference forms differed just as much in the CR skills as in the MC skills, we conducted the anchor equating by using only MC

items as the anchor. We equated the scores from the new (i.e., the second) administration to the scores from the original administration through an anchor consisting of only the MC items. The criterion equating was an anchor equating through an anchor including all the MC and CR items, with rescoring of the CR responses of the reference-form test takers by scorers from the new-form scorer pool. The reference-form test takers for this equating were approximately 500 in number, randomly selected from all the test takers tested in the original administration. Their responses were mixed with the responses of the new-form test takers and scored just like the new-form responses, with two scorers independently scoring each response and any large differences adjudicated by an expert scorer. We refer to this group as the rescored sample. Each test taker in the rescored sample thus had two CR scores, one from the scorers in the original administration and the other from scorers in the new-form administration. Because all the MC and CR items were included in the anchor, the total and anchor scores for the new-form test-taker sample (i.e., test takers from the second administration) were exactly the same. Therefore, in this special case, this anchor equating became equivalent to a single group equating in the rescored sample.

We did the MC-only anchor equating in two ways. One way was to use a reference-form equating sample that included all test takers who took the reference form, the way an MC-only anchor equating would be completed operationally. The other way was to use a reference-form equating sample that included only those test takers whose CR responses were rescored, that is, the rescored sample. We numbered these two equatings 1 and 2. The purpose of Equating 2 was to provide comparisons in which any differences from the criterion equating would be the result of using a different equating design rather than the result of using a different reference-form equating sample. See Figure 1 for a summary of the test-taker samples, scorer groups, and scores used in all of the aforementioned equatings.

To test the assumption that the CR scoring standards for new-form test takers were the same as those for reference-form test takers, we included an equating in which the anchor included all of the CR items, with no rescoring of the reference-form test takers' responses. The CR scores in the anchor score of the reference group were the scores given by the original scorers, not the new scorers. We numbered this Equating 3. Because the new form, the reference form, and the anchor for this method all included exactly the same items with no CR rescoring, the equating method in this special case is equivalent to the identity. Choosing this equating

method when an entire test form is being repeated is equivalent to assuming that the previous raw-to-scale conversion for this form is still correct.

	Equating 1: MC anchor, all ref-form test takers	Equating 2: MC anchor, rescored sample	Equating 3: MC/CR anchor, no rescoring	Equating 4: Rescoring by first scorer	Equating 5: Rescoring by second scorer	Criterion equating MC/CR-both
New-form test taker sample	all test takers from new admin.					
New-form anchor score	all MC items		all MC items + all CR (two scoring from new admin.)			
New-form total score	all MC items + all CR (two scorings from new admin.)					
Reference-form test taker sample	all test takers from previous admin.	rescored sample from previous admin.	all test takers from previous admin.	rescored sample from previous admin.		
Reference-form anchor score	all MC items		all MC items + all CR (two scorings from previous admin.)	all MC items + all CR (first scoring from rescoring at new admin., doubled)	all MC items + all CR (second scoring from rescoring at new admin., doubled)	all MC items + all CR (two scoring from rescoring at new admin.)
Reference-form total score	all MC items + all CR (two scores from previous admin.)					

Figure 1. Test-taker samples, total scores, anchor scores, and scorer groups used in equatings. MC = multiple choice; CR = constructed response.

We evaluated the effectiveness of the three anchor equatings by computing the conditional equated raw score difference from the criterion equating at each new-form score value. We plotted the conditional differences and then computed the overall weighted average bias, which we will call the root mean square difference (RMSD),

$$\text{RMSD} = \sqrt{\sum_{i=0}^{I} W_i [\hat{e}(x_i) - e(x_i)]^2} , \quad (1)$$

where i denotes a composite total raw score value; I is the maximum possible total score, 113 or 114, in the test; $\hat{e}(x_i)$ is the reference-form equivalent of new-form score value x_i calculated from the anchor equating; and $e(x_i)$ is the reference-form equivalent of new-form score value x_i from the criterion equating. W_i is the proportion of the new-form test takers at score point x_i.

We also calculated the bootstrap conditional equating bias and conditional equating error for each of the three anchor equatings. This calculation was done through replicating each anchor equating 500 times using bootstrap samples, which were obtained by resampling (with replacement) from each of the new-form and reference-form groups. The bootstrap conditional equating bias at each raw score value is the difference between the anchor equating and the criterion equating, averaged over 500 replications. It is a measure of systematic error. The bootstrap conditional standard error of equating (SEE) at each raw score value is the standard deviation of the equated score over 500 replications. It is a measure of random sampling error in equating. The sum of the SEE and squared bias represents the total equating error, and the square root of this value is the conditional root-mean-squared error (RMSE), which represents the combination of systematic and random errors. The equations for bootstrap conditional equating bias, SEE, and RMSE are as follows:

$$\text{bias}(x_i) = \frac{1}{R}\sum_{r=1}^{R}[\hat{e}_r(x_i)] - e(x_i)] \ , \tag{2}$$

$$\text{SEE}(x_i) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}[\hat{e}_r(x_i) - \overline{\hat{e}}(x_i)]^2} \ , \tag{3}$$

$$\text{RMSE}(x_i) = \sqrt{[\text{bias}(x_i)]^2 + [\text{SEE}(x_i)]^2} \ , \tag{4}$$

where r denotes a replication and R is the total number of replications (equal to 500 in this study); $\hat{e}_r(x_i)$ is the reference-form equivalent of new-form score point x_i from the anchor equating in the rth replication, and $\overline{\hat{e}}(x_i)$ is the average of $\hat{e}_r(x_i)$ over the R replications. We averaged equating bias, conditional SEE, and conditional RMSE over the raw total score frequencies of the new-form group. This evaluation produced three single statistics based on the bootstrap estimates: the weighted average bias, the weighted average SEE, and the weighted average RMSE. We named the weighted average bias the *bootstrap weighted average bias* to distinguish it from RMSD, which is a direct estimate of the weighted average bias. The equations for these three indices are as follows:

$$\text{bootstrap weighted average bias} = \sqrt{\sum_{i=0}^{I} W_i[\text{bias}(x_i)]^2} \ , \tag{5}$$

$$\text{weighted average SEE} = \sqrt{\sum_{i=0}^{I} W_i [\text{SEE}(x_i)]^2}, \quad (6)$$

$$\text{weighted average RMSE} = \sqrt{\sum_{i=0}^{I} W_i [\text{RMSE}(x_i)]^2}, \quad (7)$$

where i denotes a composite total raw score value, I is the number of maximum possible score levels (114 in this test), and W_i is the proportion of the new-form test takers at total raw score point i.

For each equating, we also computed coefficient kappa and the percentage of agreement (i.e., the percentage of the test takers who would be classified the same way, pass or fail, as they were by the criterion equating). Because different test users use different cut scores, we selected three different cut scores, the lowest, highest, and most frequent, to conduct this analysis.

Analysis, Part II

To evaluate how closely the results of the partial rescoring matched the results of full rescoring, we conducted two more anchor equatings. The anchor scores of the reference-form test takers were based on rescoring by the new-form scorers, but only one score of each response was used in computing the anchor score. That score was multiplied by 2 to give the item its appropriate weight in the total score. The total score of the reference-form test takers was based on the operational scoring of the original administration, in which each response was scored by two independent scorers with adjudication. Both the total and anchor scores of the new-form test takers were based on the operational scoring. In this study, the anchor included all the items, both MC and CR. With the same items and the same scoring procedure, the anchor scores of the new-form test takers were exactly the same as their total scores. Therefore, each of these anchor equatings became identical to a single group equating based on the rescored sample. We numbered these two equatings 4 and 5, with Equating 4 based on the first score of each response in the rescoring and Equating 5 based on the second. The fourth and fifth columns of Figure 1 show test-taker samples, scorer groups, and scores used in these two equatings. Each equating was compared to the criterion equating used in Part I, which employed two scorers. As in Part I, the statistics based on Equations 1 through 7, percentage agreement, and coefficient kappa were calculated for Equatings 4 and 5.

The chained equipercentile equating method was used for Equatings 1 and 2. The direct equipercentile method was used for Equatings 4 and 5 and the criterion equating. Data were presmoothed using the log-linear smoothing method (Holland & Thayer, 1987) with four moments.

Evaluation Criterion

The number of raw score points for this test was about equivalent to the number of scale score points. When evaluating an equating result operationally, a difference in equating methods that is greater than one half of a raw score point, which is about 5% of the standard deviation of the raw total score, is typically considered large for this test because this difference tends to lead to a change in the scale score.

Results

Part I

Figures 2 and 3 show how Equatings 1, 2, and 3 differ from the criterion equating at each new-form raw score level for Forms A and B, respectively. Equatings 1 and 2 used the MC anchor; Equating 3 used the MC/CR anchor without rescoring. The conditional differences for new-form total raw scores of 50 and below are not shown because almost no test takers scored below 50.

The equatings that used an MC/CR anchor without rescoring of the CR responses (i.e., Equating 3) differed substantially from the criterion. The discrepancies indicated an existence of a scoring shift across the administrations, and the patterns of the scoring shift for the two forms appeared to be different. For Form A, the equated scores were below the criterion in the middle of the scale, indicating that the new scorer group was more stringent for responses of average quality. The largest difference in the middle was slightly over 1 raw score point. For Form B, the differences increased as test takers' total scores decreased, indicating that the new scorer group was more lenient in scoring responses of below-average quality. The differences at the lower end of the raw scale were more than 1.5 raw score points. With no rescoring, there was no adjustment for the shift in scoring standards.

The two equatings that used the MC-only anchor produced similar results in the middle of the raw score scale but deviated from each other at the high and low ends of the scale. If we focus on the area of the scale in which the middle 80% of the test takers' scores (roughly between raw score points 75 and 90), most of the differences were within .50 raw score points.

The differences gradually became larger beyond this area of the scale. In the middle of the scale, these two equatings were closer to the criterion than the equating that used the MC/CR anchor without rescoring. However, beyond the 10th and 90th percentiles, the differences for these equatings were much larger than the equating that used the MC/CR anchor without rescoring.

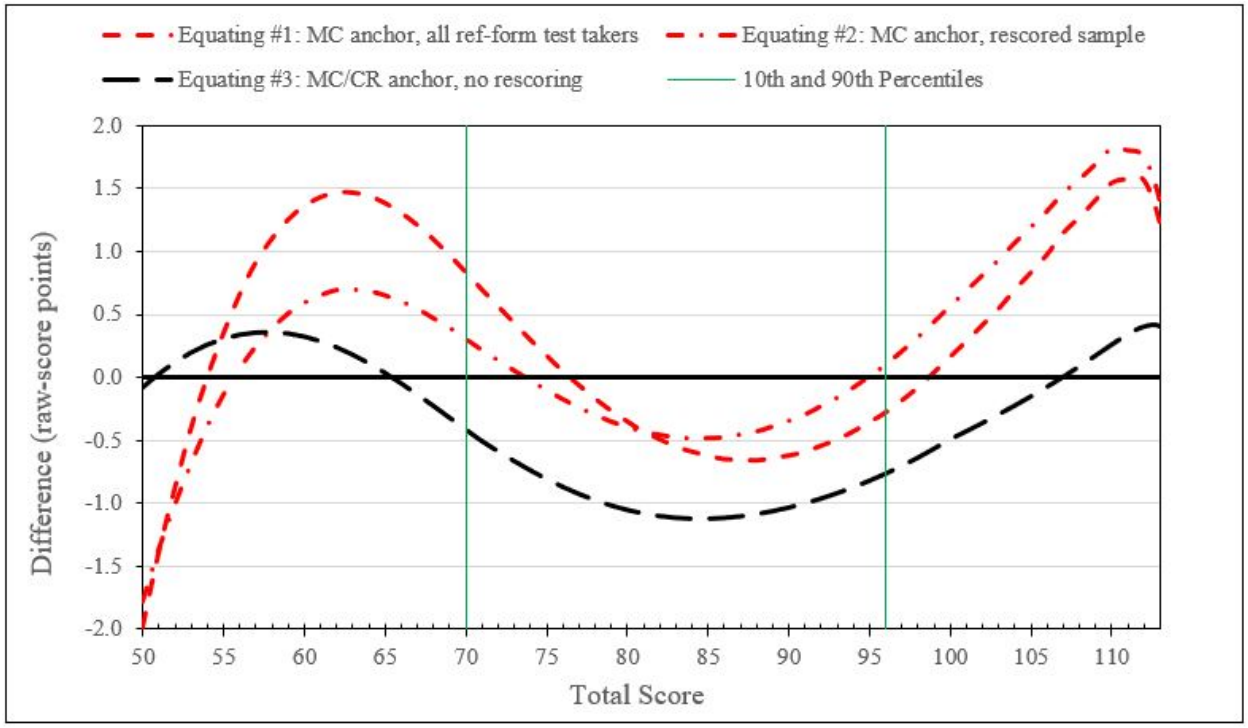

Figure 2. Form A: Difference from criterion equating, for equatings through multiple-choice-only anchor and multiple-choice/constructed-response anchor without rescoring.

The RMSD, bootstrap weighted average bias, SEE, and RMSE from Equatings 1, 2, and 3 for Forms A and B are presented in Table 1. For Form A, the equating that used the rescored sample yielded a smaller RMSD (.58) than the equating that used the whole reference group sample (.80). For Form B, the equating that used the rescored sample yielded a larger RMSD (.84) than the equating that used the whole reference group sample (.68) and the MC/CR equating without rescoring CR (.75). However, as Figure 3 shows, if we had calculated the RMSD only for the middle 80% of the test takers, the two MC-only anchor equatings would have yielded much smaller RMSD values, similar for the two forms. For Equatings 1, 2, and 3, the RMSD values would be .50, .34, and .98 for Form A and .50, .44, and .68 for Form B, respectively.

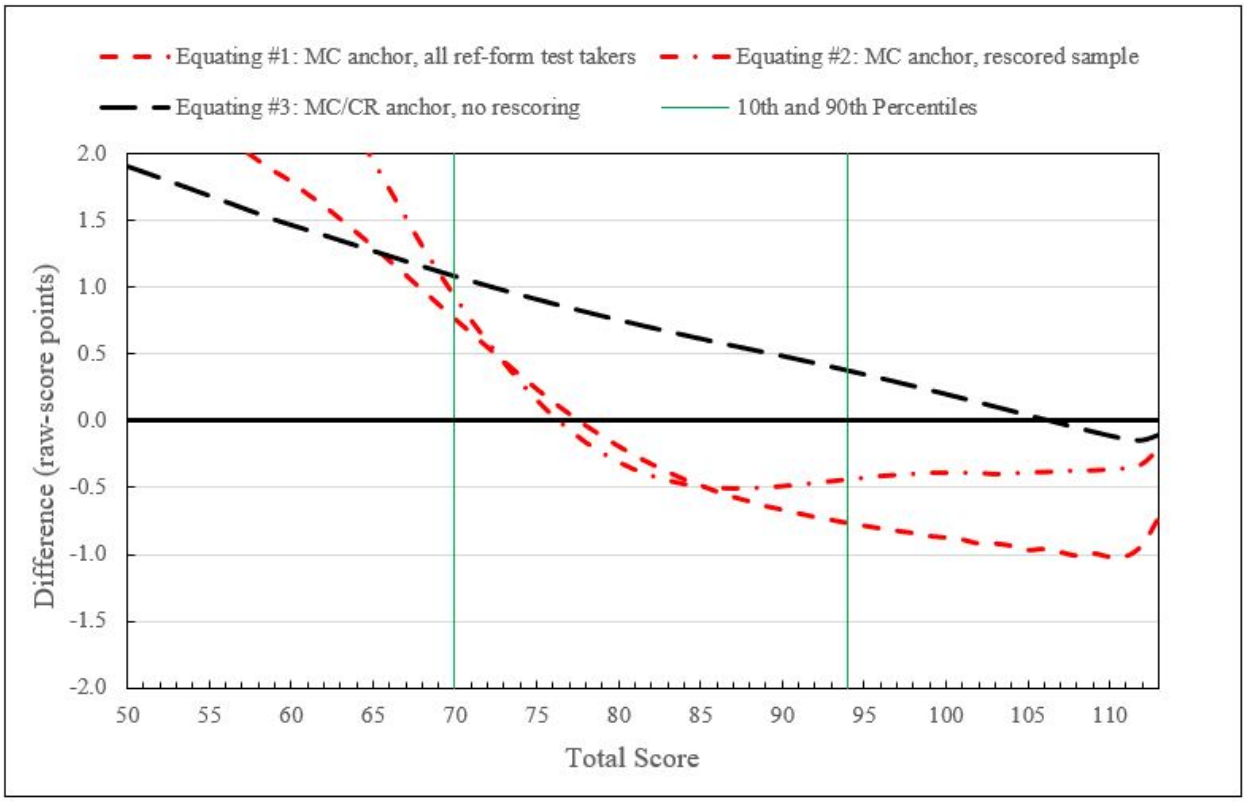

Figure 3. Form B: Difference from criterion equating, for equatings through multiple-choice-only anchor and multiple-choice/constructed-response anchor without rescoring.

Table 1. Summary of Root-Mean-Square Difference, Weighted Average Bias, Weighted Average Standard Error of Equating, and Weighted Average Root-Mean-Square Error Between Alternative Anchor Equatings and Criterion Equating

Anchor equatings	*RMSD*	Bootstrap weighted average bias	Weighted average *SEE*	Weighted average *RMSE*
Form A				
Equating 1: MC anchor, all ref.-form test takers	.80	.77	.40	.86
Equating 2: MC anchor, rescored ref.-form sample	.58	.56	.49	.74
Equating 3: MC/CR anchor, no rescoring[a]	.91	.91	.00	.91
Form B				
Equating 1: MC anchor, all ref.-form test takers	.68	.67	.28	.73
Equating 2: MC anchor, rescored ref.-form sample	.84	.84	.41	.93
Equating 3: MC/CR anchor, no rescoring[a]	.75	.75	.00	.75

Note. CR = constructed response; MC = multiple choice.
[a]Equal to the identity in this special case of a repeated test form.

The bootstrap weighted average biases are consistent with the RMSD. Note that the equating without rescoring the CR items is equivalent to the identity, which is not affected by sampling variability; consequently, the SEE was zero. The overall equating errors, as summarized in RMSE, for the MC-only anchor and MC/CR anchor without rescoring CR were all greater than .50 raw score point. The plots of the conditional equating bias for Equatings 1 and 2, along with a 68% equating error band over the bootstrap samples, are presented in Figures A3 and A4 for Form A and Figures A5 and A6 for Form B, respectively.

The pass–fail percentage agreement and coefficient kappa between the criterion and each of Equatings 1, 2, and 3 at the lowest, most frequently used, and highest cut scores are presented in Table 2. The raw scores for the lowest cut score of the criterion equatings were 65 and 63 for Form A and Form B, respectively, about the 5th percentile of score distribution. The raw scores for the most frequent cut scores of the criterion equating were 75 and 73 for Form A and Form B, respectively, which were equivalent to the 15th percentile. The statistics in Table 2 are all very high, at either 100% agreement or just slightly lower. In those cases, the difference in the two raw scores was only 1 point, except for the case of Equating 2 in Form B. For that particular equating, the difference was 2 raw score points, and the coefficient kappa was .75, the lowest among all. Both Equatings 1 and 2 yielded a perfect agreement with the criterion equating at the most popular cut score for Forms A and B.

Table 2. Percentage Agreement (Coefficient Kappa) Between the Criterion and Alternative Equating Functions at Three Cut Scores

Anchor equating	Lowest cut score	Highest cut score	Most frequent cut score
Form A			
Equating 1: MC anchor, all ref.-form test takers	99% (.93)	100% (1.0)	100% (1.0)
Equating 2: MC anchor, rescored sample	100% (1.0)	97% (.93)	100% (1.0)
Equating 3: MC/CR anchor, no rescoring[a]	100% (1.0)	97% (.93)	98% (.93)
Form B			
Equating 1: MC anchor, all ref.-form test takers	98% (.86)	100% (1.0)	100% (1.0)
Equating 2: MC anchor, rescored sample	97% (.75)	100% (1.0)	100% (1.0)
Equating 3: MC/CR anchor, no rescoring[a]	98% (.86)	98% (.92)	97% (.91)

Note. CR = constructed response; MC = multiple choice.
[a]Equal to the identity in this special case of a repeated test form

The summary statistics of total and anchor scores on the new and reference forms for the equatings conducted in Part I of the study for Forms A and B are presented in Table 3. In the criterion equating of Form A, the total and anchor score means on the reference form were 85.2

and 84.3 (note that the maximum possible raw score is 114). The total and anchor scores were both based on all the items in the test. The difference of these two means reflected the scoring shift in Form A, which was −0.84. Similarly, in Form B, the scoring shift was 0.72. Note in Table 3 that if two numbers are exactly the same (e.g., the total and anchor score means of the new form in the criterion equating), it is because they refer to the same statistics computed from the same data. The correlation between the total and anchor scores for Equatings 1 and 2 across the new and reference groups of the two forms ranged from .88 to .90.

Table 3. Summary Statistics of the New- and Reference-Form Total and Anchor Scores Used in Equatings 1, 2, and 3 for Forms A and B

Anchor equating	Sample size new form	Total mean (*SD*) new form	Anchor mean (*SD*) new form	Sample size reference form	Total mean (*SD*) reference form	Anchor mean (*SD*) reference form
Form A						
Criterion	2,053	83.8 (10.2)	83.8 (10.2)	532	85.2 (10.6)	84.3 (10.4)
Equating 1: MC anchor, all ref.-form test takers	2,053	83.8 (10.2)	62.2 (6.1)	1,342	86.1 (9.7)	63.2 (5.8)
Equating 2: MC anchor, rescored ref.-form sample	2,053	83.8 (10.2)	62.2 (6.1)	532	85.2 (10.6)	62.6 (6.2)
Equating 3: MC/CR anchor, no rescoring[a]	2,053	83.8 (10.2)	83.8 (10.2)	532	85.2 (10.6)	85.2 (10.6)
Form B						
Criterion	1,562	82.9 (9.4)	82.9 (9.4)	547	81.2 (9.5)	82.0 (9.2)
Equating 1: MC anchor, all ref.-form test takers	1,562	82.9 (9.4)	59.3 (5.8)	2,045	81.2 (9.6)	58.8 (6.2)
Equating 2: MC anchor, rescored ref.-form sample	1,562	82.9 (9.4)	59.3 (5.8)	547	81.2 (9.5)	58.7 (6.2)
Equating 3: MC/CR anchor, no rescoring[a]	1,562	82.9 (9.4)	82.9 (9.4)	547	81.2 (9.5)	81.2 (9.5)

Note. CR = constructed response; MC = multiple choice.
[a]Equal to the identity in this special case of a repeated test form.

Part II

The plots of the equated total score difference for equatings based on a single-scorer rescoring of CR responses (i.e., Equatings 4 and 5) are presented in Figure 4 and Figure 5 for Form A and Form B, respectively. In each plot, both curves are close to the zero line. However, for low scorers, all four curves are above the zero line; for high scorers, three of the four curves are below the zero line. This difference indicates that the equatings based on rescoring of CR responses by a single scorer produced new-form conversions that were too high for low-scoring test takers and too low for high-scoring test takers. In Form B, how could Equatings 4 and 5

differ in the same direction from the criterion equating? There is only one possibility: The criterion equating included adjudication of differences between the two scorers doing the rescoring, but Equatings 4 and 5, with only one score, could not include adjudication.

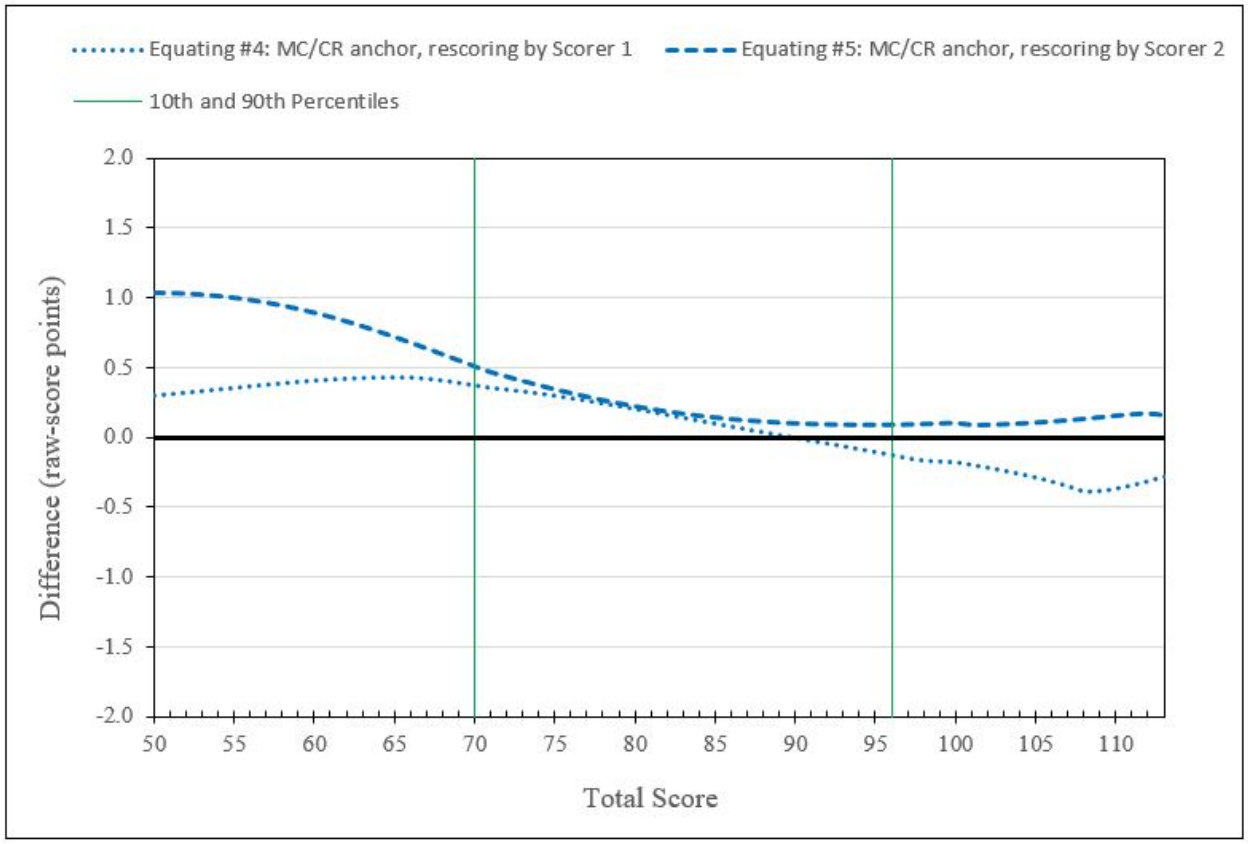

Figure 4. Form A: Difference from criterion equating, for equating based on single-scorer rescoring.

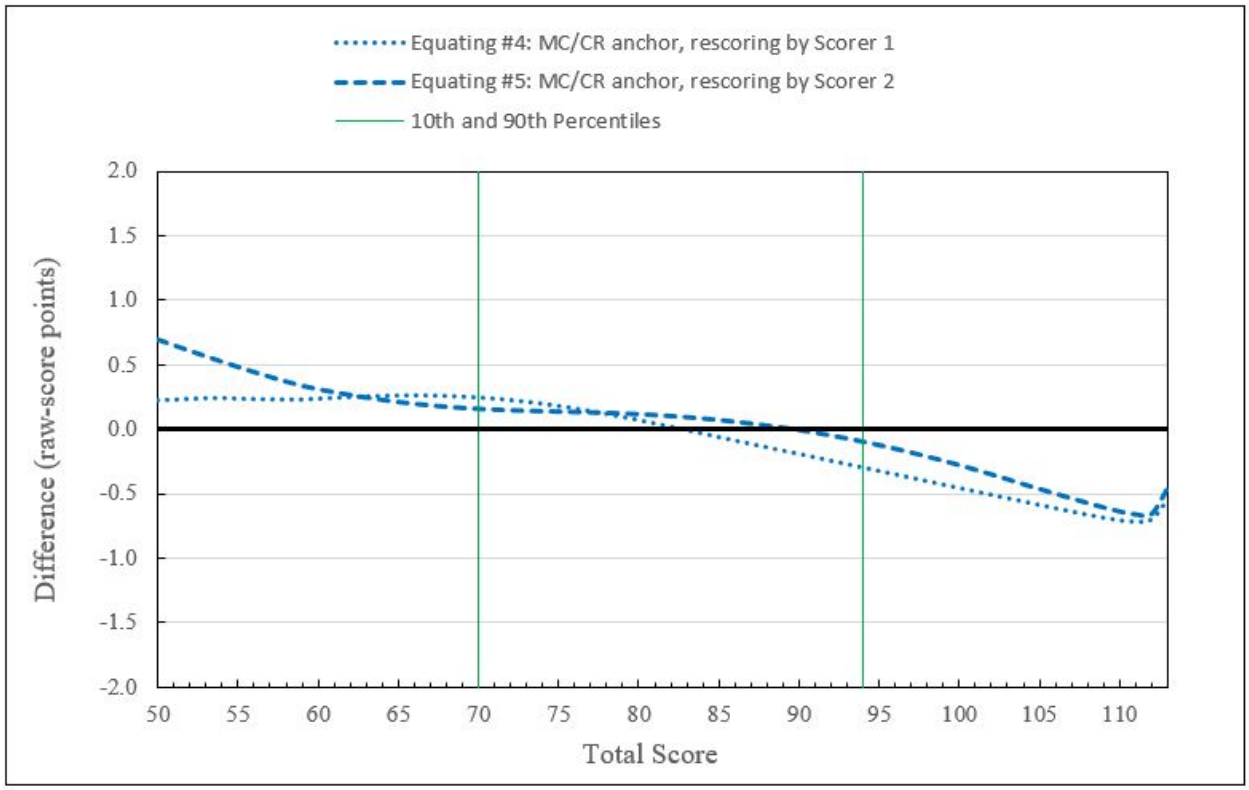

Figure 5. Form B: Difference from criterion equating, for equating based on single-scorer rescoring.

The RMSD, bootstrap weighted average bias, SEE, and RMSE for the two equatings that used a single scorer to rescore CR responses (i.e., Equatings 4 and 5) are presented in Table 4. As Figures 4 and 5 show, these equatings yielded very small RMSD values, ranging from .13 to .30. The overall equating errors, as indicated by the RMSE value, were below .50 for both Forms A and B. The plots of the conditional biases for Equatings 4 and 5, along with a 68% equating error band based on the bootstrap samples, are presented in Figures A7 and A8 for Form A and Figures A9 and A10 for Form B, respectively.

Table 4. Summary of Root-Mean-Square Difference, Weighted Average Bias, Weighted Average Standard Error of Equating, and Weighted Average Root-Mean-Square Error Between One-Scorer Rescoring and Two-Scorer Rescoring Equating

Anchor equating	*RMSD*	Bootstrap weighted average bias	Weighted average *SEE*	Weighted average *RMSE*
Form A				
Equating 4: MC/CR anchor, rescoring by first scorer	.20	.21	.25	.33
Equating 5: MC/CR anchor, rescoring by second scorer	.30	.30	.28	.41
Form B				
Equating 4: MC/CR anchor, rescoring by first scorer	.20	.20	.24	.31
Equating 5: MC/CR anchor, rescoring by second scorer	.13	.13	.24	.27

Note. CR = constructed response; MC = multiple choice.

The pass–fail decision percentage agreement and coefficient kappa at the three selected cut scores were all 100%, except for Equating 5 at the lowest cut score. See Table 5.

Table 5. Agreement Percentages (Coefficient Kappa) Between Pass/Fail Decisions Based on One-Scorer Rescoring and Two-Scorer Rescoring

Anchor equating	Low cut score	High cut score	Most frequent cut score
Form A			
Equating 4: MC/CR anchor, rescoring by first scorer	100% (1.0)	100% (1.0)	100% (1.0)
Equating 5: MC/CR anchor, rescoring by second scorer	99% (.93)	100% (1.0)	100% (1.0)
Form B			
Equating 4: MC/CR anchor, rescoring by first scorer	100% (1.0)	100% (1.0)	100% (1.0)
Equating 5: MC/CR anchor, rescoring by second scorer	100% (1.0)	100% (1.0)	100% (1.0)

Note. CR = constructed response; MC = multiple choice.

The summary statistics and reliability coefficients of total scores used in Equatings 4 and 5 for Forms A and B are presented in Table 6. These two equatings were single group equatings based on the rescored samples. The total score means based on the rescored scorings from Scorers 1 and 2 were comparable in Form A (84.2 vs. 84.1) and Form B (82.0 vs. 81.9). However, it should be pointed out that the averaging process makes the two effects of the equatings (i.e., a new-form conversion that was too high for low-scoring test takers and too low for high-scoring test takers) cancel each other out. The correlations of the total scores based on

Scorer 1 and Scorer 2 for Forms A and B were approximately .94. The reliability of the total scores based on double scoring was .81 or .82 for Form A and .77 or .78 for Form B. When switching to the single-scorer model, the reliability was only reduced by .02 and .04 for Form A and Form B, respectively. The reliability of the total score was a stratified alpha, estimated based on the coefficient alphas of the MC and CR scores.

Table 6. Summary Statistics and the Reliability of the Raw Total Scores Used in Single Group Equatings 4 and 5 for Forms A and B

Anchor equating	Sample size	Rescored *M* (*SD*) [reliability]	Original *M* (*SD*) [reliability]
Form A			
Criterion equating	532	84.3 (10.4) [0.81]	85.2 (10.6) [0.82]
Equating 4: rescoring by first scorer	532	84.2 (10.6) [0.79]	85.2 (10.6) [0.82]
Equating 5: rescoring by second scorer	532	84.1 (10.6) [0.79]	
Form B			
Criterion equating	547	82.0 (9.2) [0.78]	81.2 (9.5) [0.77]
Equating 4: rescoring by first scorer	547	82.0 (9.4) [0.74]	81.2 (9.5) [0.77]
Equating 5: rescoring by second scorer	547	81.9 (9.3) [0.74]	

Discussion

The first part of the study focused on examining whether using a MC-only anchor or an MC/CR anchor without rescoring the CR responses would work appropriately for a mixed-format test in which the MC section makes up a very large portion of the total score. We investigated the problem for a licensure test in which the MC section makes up 70% of the total score, and the disattenuated correlation of MC and CR section scores was about .70. The two anchor designs were the same as those used in Kim et al. (2010), but in their study, the MC items accounted for only one-third of the total score, and the anchor consisted of only approximately half of the MC items—approximately one-sixth of the total score. In their study focusing on equating two different test forms, they found that the overall equating errors (i.e., the RMSE) associated with these two types of anchors were approximately equal and approximately 1.5 raw score points. The test forms to equate in the current study actually were the same form, which allowed us to use the whole MC, 70% of the total score, as the anchor. In practice, the repeated MC items in a new form of this test (and therefore available for use as an equating anchor) typically account for no more than 35% of the total score. In this study, with an MC anchor accounting for 70% of the total score, the overall equating errors were still more than 0.5 raw score points. In our study, we found that the MC-only anchor equating could be more accurate

than the MC/CR anchor equating without rescoring CR items, but only for test takers who scored between the 10th and 90th percentiles. The anchor equating using the MC-only anchor tends to yield larger differences than the equating using the MC/CR anchor without rescoring of the CR, for higher and lower scorers. In this study, we also found that the two MC-only anchor equatings provided similar results regardless of whether the rescored sample or the total reference-form sample was used as the reference-form group in the equating.

In the second part of the study, we investigated whether using a single scorer in the rescoring process could replace the full-scoring process. The results showed that using a single scorer in the rescoring process could approximate the results of using two scorers for the majority of test takers. However, there seemed to be an adjudication effect. Consider the difference at the higher end of the score scale on Form B, in which the equated new-form scores from Equatings 4 and 5 are both too low. The new-form data and analysis were exactly the same in Equatings 4 and 5 as in the criterion equating, so the cause must be in the reference-form rescoring. But why would using only one of the two scores from the rescoring tend to underestimate the ability of the strongest test takers? That would happen if there was a tendency for the (relatively) low scores of these strong test takers' responses in the rescoring not to be used in the criterion equating. At the low end of the score scale, the effect is exactly the opposite. The scoring not used in the criterion equating (because they were replaced as a result of adjudication) tended to be the high scores. Using them in Equatings 4 and 5 made the reference form appear easier than it really was for test takers of low ability. As a result, Equatings 4 and 5 were too high at the low end of the score scale. However, this effect might have disappeared if we had conducted Equatings 4 and 5 using an anchor design, with the anchor scores of both new-form and reference-form test takers based on a single scoring of the CR responses. In that case, there would be no adjudication affecting the scores of either group. Future studies could be conducted using an anchor design to equate a new or reprint form, to check whether the effect found in this study would disappear.

The pass–fail percentage agreement and coefficient kappa were in general very high. There were more inconsistencies at the lowest cut scores. This is because there were very few people in the area of the lowest cut score, so equating results tended not to be stable. The equatings that used the MC-only anchor and rescoring of the reference-form CR responses by one scorer yielded consistent results at the most frequent cut scores and at the highest cut scores.

These results implied that equatings with the MC-only anchor design and the one-scorer rescoring design may be good when the cut scores are in the middle of the score scale. However, even though a single small inaccuracy may not affect pass–fail decisions, a chain of scoring shifts in the same direction could result in a substantial change in the effective standards for passing the test. Also, for some testing programs, different users of the test have different cut scores. For convenience, we analyzed only three selected cut scores for illustration. We think it is important that a testing program in which cut scores would change should evaluate the score comparability along the whole score scale, not just on certain cut score points.

Conclusions and Limitations

In the first part of this study, we found that the MC/CR anchor without rescoring of the reference-form CR anchor item responses yielded poor equating results, indicating that there had been a scoring shift. If a more severe scoring shift had been observed, the results could have been worse. This result indicates that if reference-form CR anchor item responses cannot be rescored in equating, it would be better not to include CR items in the anchor. In operation, no matter how hard a testing program tries to maintain the same scoring standards across time, a scoring shift can still occur. The MC-only anchor results found in this study were more promising, but not completely satisfactory. The fact that using an anchor that made up 70% of the items in the test (i.e., the MC items) could not completely achieve a satisfactory result indicates that the groups did not vary exactly in the same way in their MC skills as in their CR skills. It is not clear whether a more satisfactory answer would be achieved if the MC section score were to make up a larger proportion of the total score or the MC and CR correlation were higher. To achieve a more definite answer to the question, we believe that future studies should use simulation data for analyses. Simulation data would allow for flexibly manipulating the weight of the MC section in the total score and the MC/CR correlation based on different test forms to be equated. Because this study was based on real data from only one test, the findings may only be generalized to other tests with similar MC/CR proportions and correlations.

In the second part of the study, we found that the equating based on the rescoring by a single scorer approximated the equating using the full-scoring procedure quite well. However, we also found an issue related to this method. Because a single scoring does not provide a way to identify possibly incorrect scorings for adjudication, the new equating tended to yield lower equated scores for strong test takers and higher equated scores for test takers of weaker ability.

When scorers disagree about the quality of a high-scoring test taker's response to a CR item, it is more likely that the higher scoring will be correct, as determined by adjudication. And similarly, the opposite will happen at the low end of the scale. Future studies exploring the possibility of reducing CR scoring to a single scorer should investigate whether this effect exists if the equating uses one score of each response to compute the anchor score of the new-form test takers.

References

Bock, R. D. (1995). Open-ended exercises in large-scale educational assessment. In L. B. Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 305–338). San Francisco, CA: Jossey-Bass

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. https://doi.org/10.1207/s15324818ame1102_5

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (PSR Technical Report No. 87-79; Research Report No. RR-87-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00235.x

Kim, S., & Moses, T. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing*, *13*, 314–328. https://doi.org/10.1080/15305058.2013.776050

Kim, S., Walker, M. E., & McHale, F. (2010). Comparison among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, *47*, 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2, CASMA Monograph No. 2.2). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336–346. https://doi.org/10.1111/j.1745-3984.1999.tb00560.x

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. https://doi.org/10.1111/j.1745-3984.2000.tb01090.x

Walker, M. E., & Kim, S. (2010). *Examining two strategies to link mixed-format tests using multiple-choice anchors* (Research Report No. RR-10-18). Princeton, NJ: Educational Testing Service.

Wang, W. (2013). *Mixed-format test score equating: Effect of item-type multidimensionality, length and composition of common-item set, and group ability difference* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Appendix

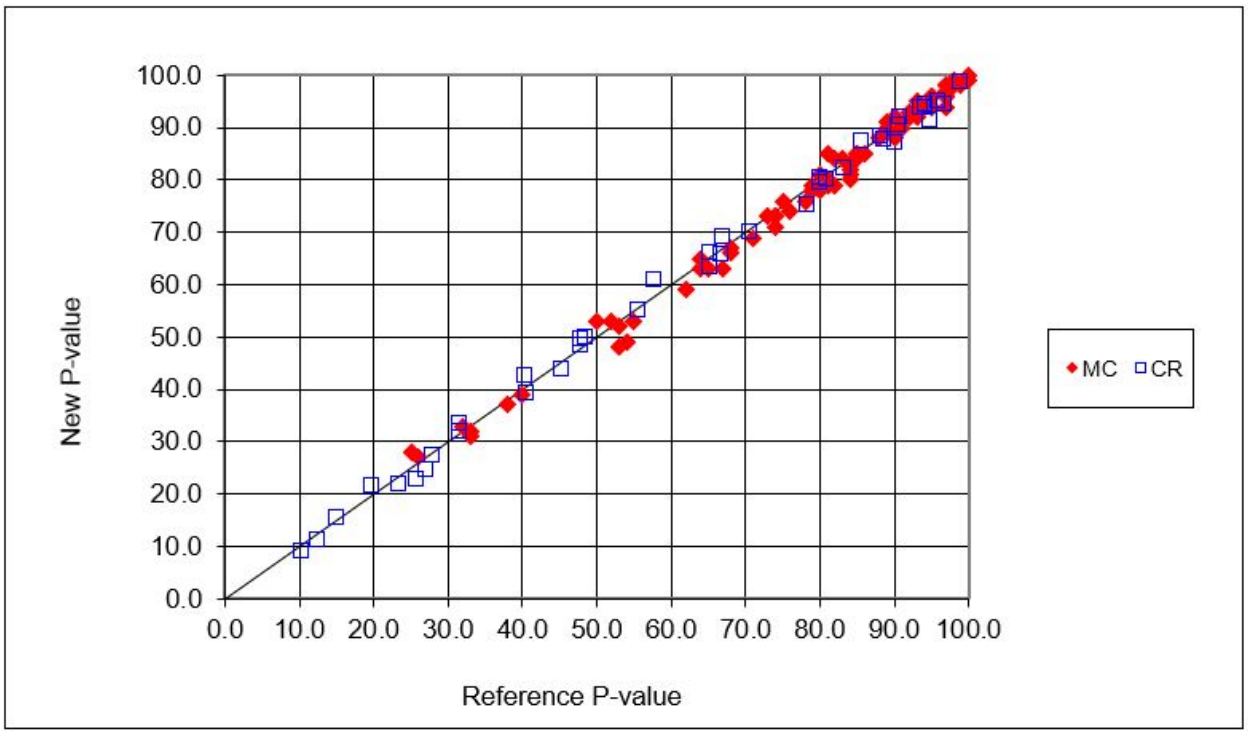

Figure A1. Form A multiple-choice and constructed-response item *p*-value plot.

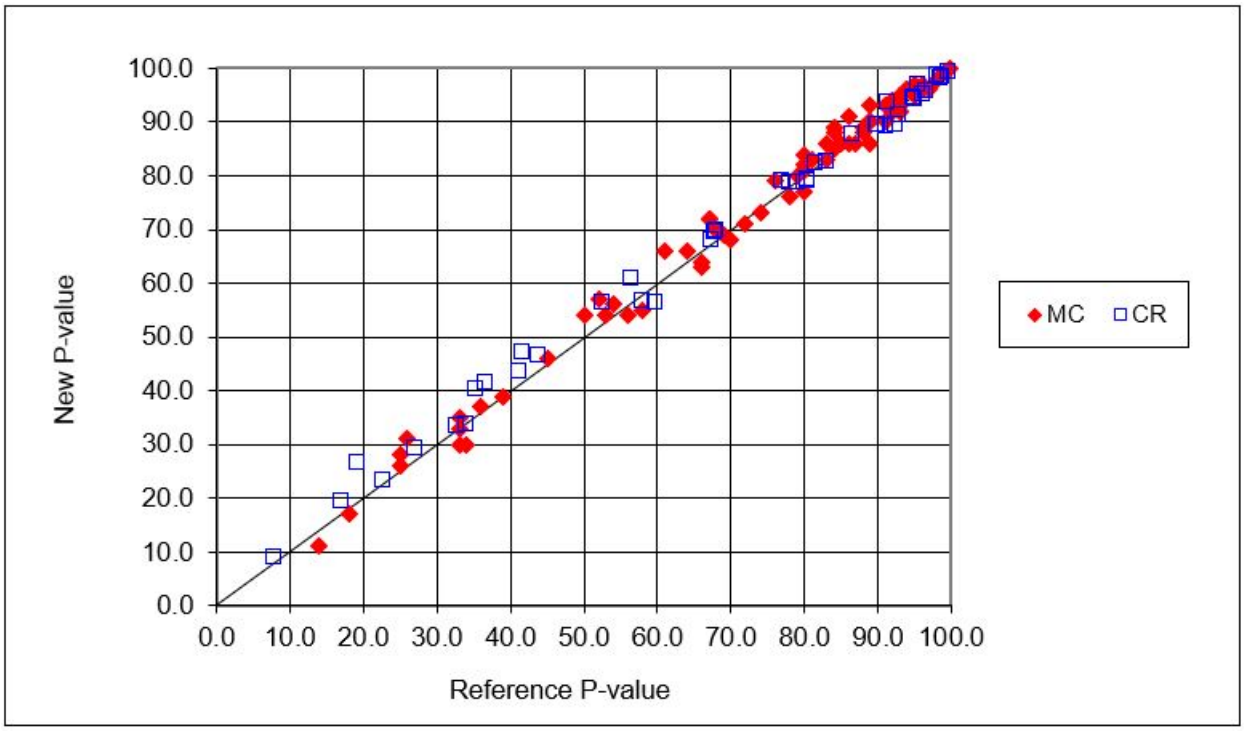

Figure A2. Form B multiple-choice and constructed-response item p-value plot.

Table A1. Correlation of *p*-Values in New Form and Reference Form Groups

Item	Form A	Form B
MC	.986	.993
CR	.997	.996
All	.992	.994

Note. CR = constructed response; MC = multiple choice.

Table A2. Means and Standard Deviations of *p*-Values

Measurement	Form A New	Form A Reference	Form B New	Form B Reference
N	2,053	1,342	1,562	2,045
MC (std)	.75 (.23)	.74 (.23)	.78 (.19)	.78 (.19)
CR (std)	.69 (.27)	.68 (.28)	.62 (.28)	.63 (.28)
All (std)	.73 (.24)	.72 (.25)	.73 (.24)	.73 (.24)

Note. CR = constructed response; MC = multiple choice.

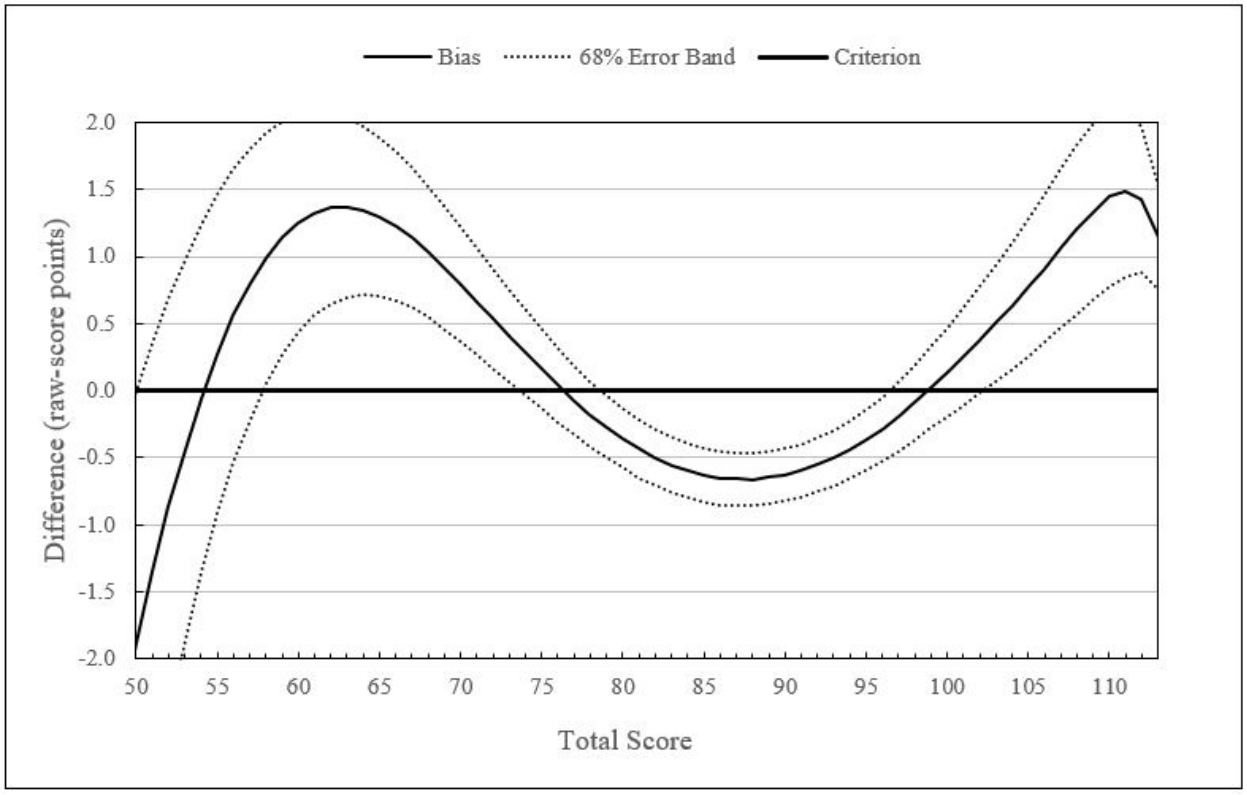

Figure A3. Form A: multiple-choice anchor, all ref-form test takers (Equating 1).

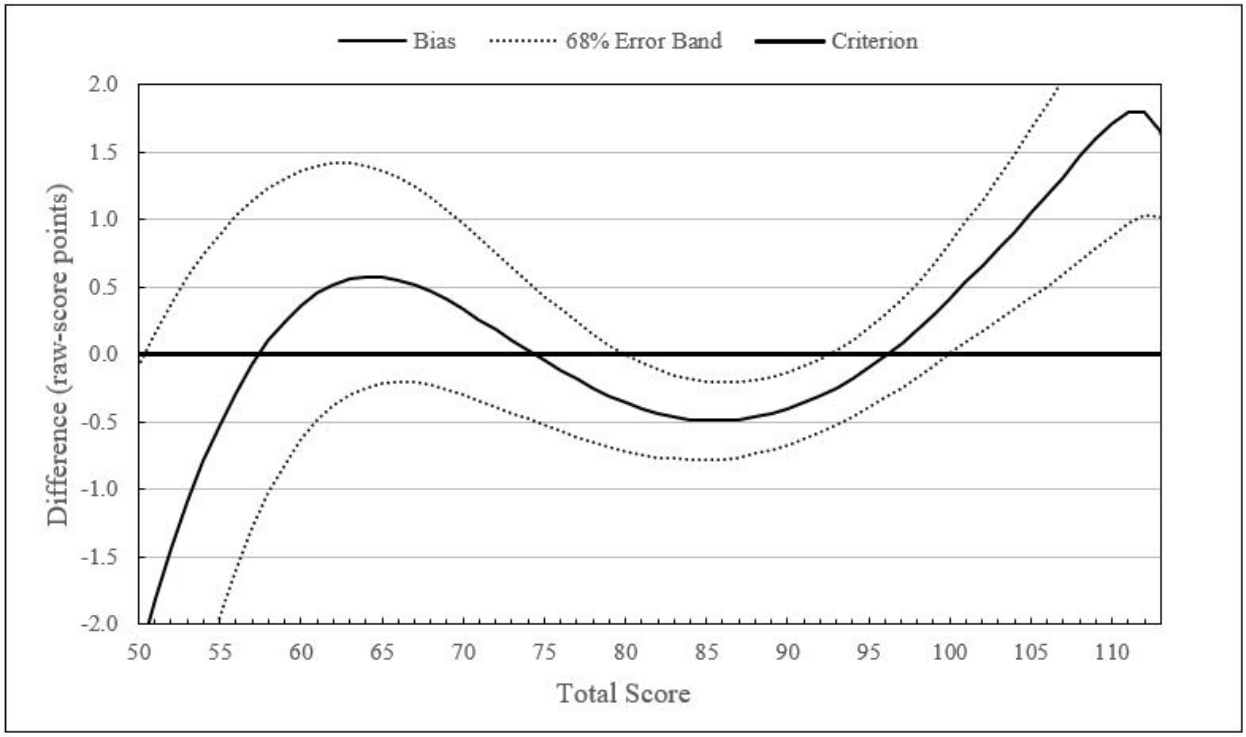

Figure A4. Form A: multiple-choice anchor, rescored ref-form sample (Equating 2).

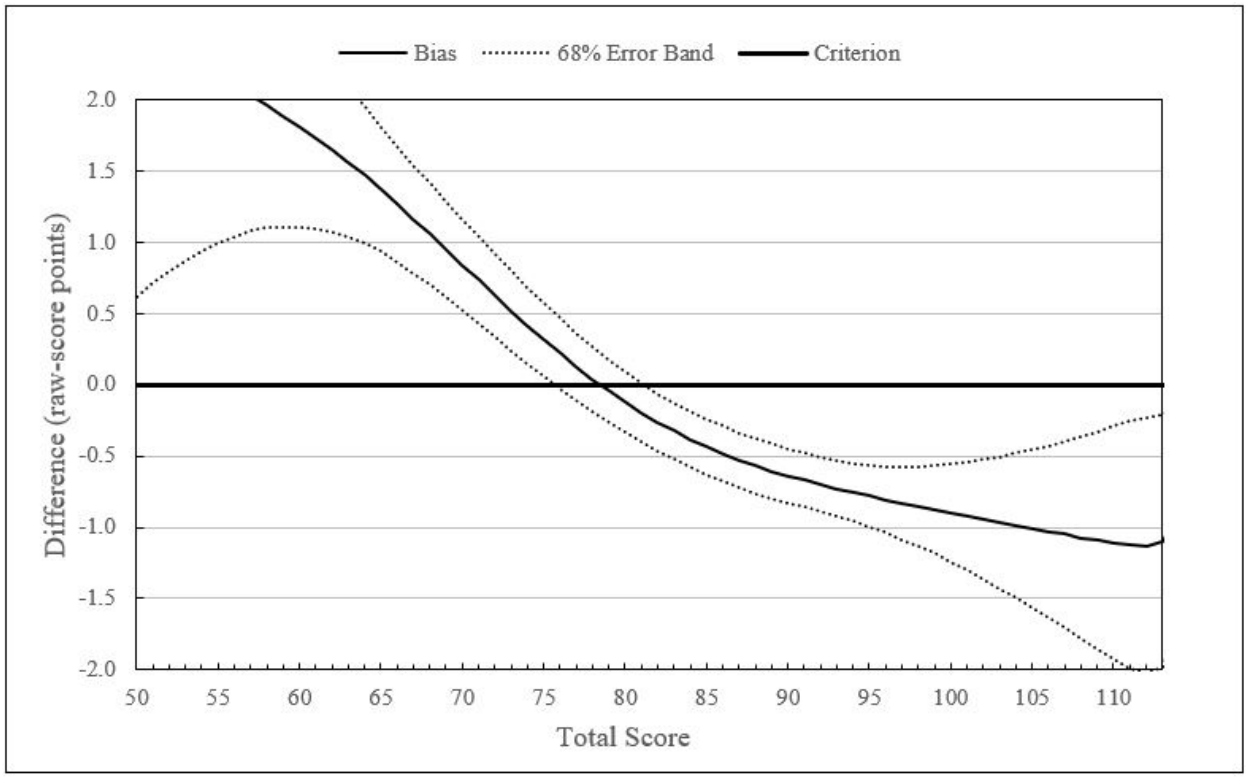

Figure A5. Form B: multiple-choice anchor, all ref-form test takers (Equating 1).

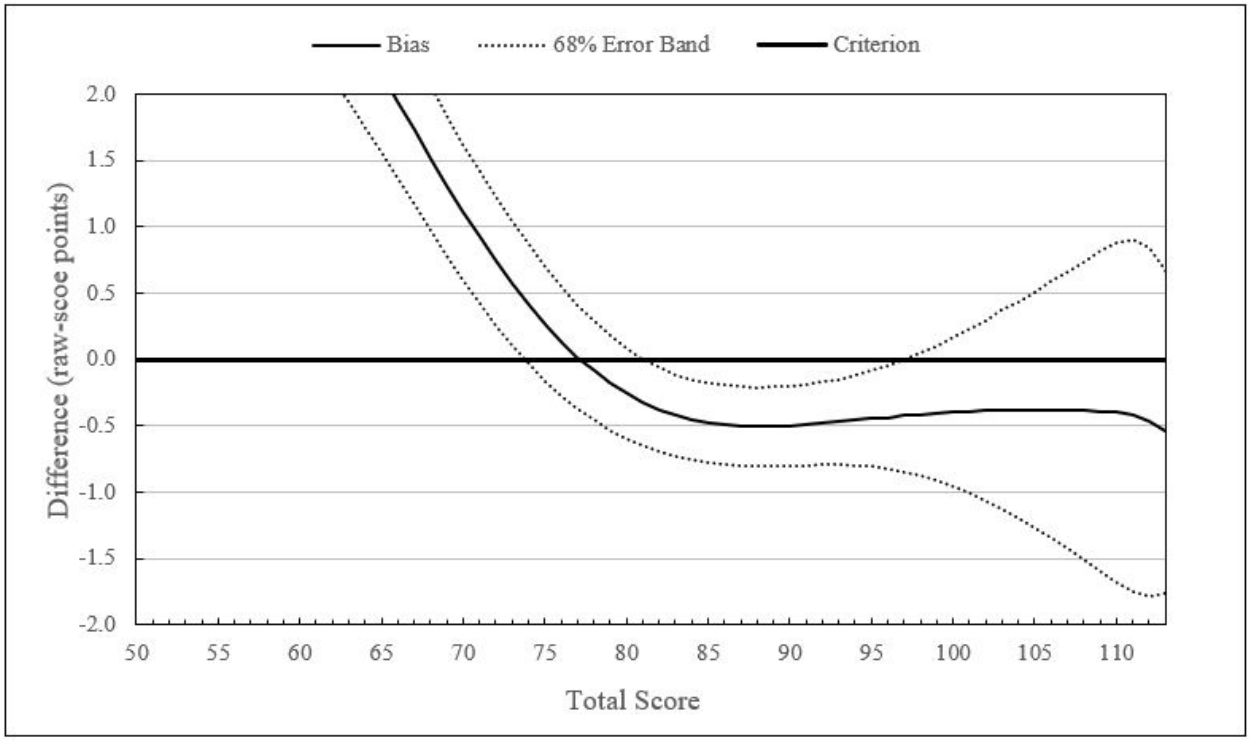

Figure A6. Form B: multiple-choice anchor, rescored ref-form sample (Equating 2).

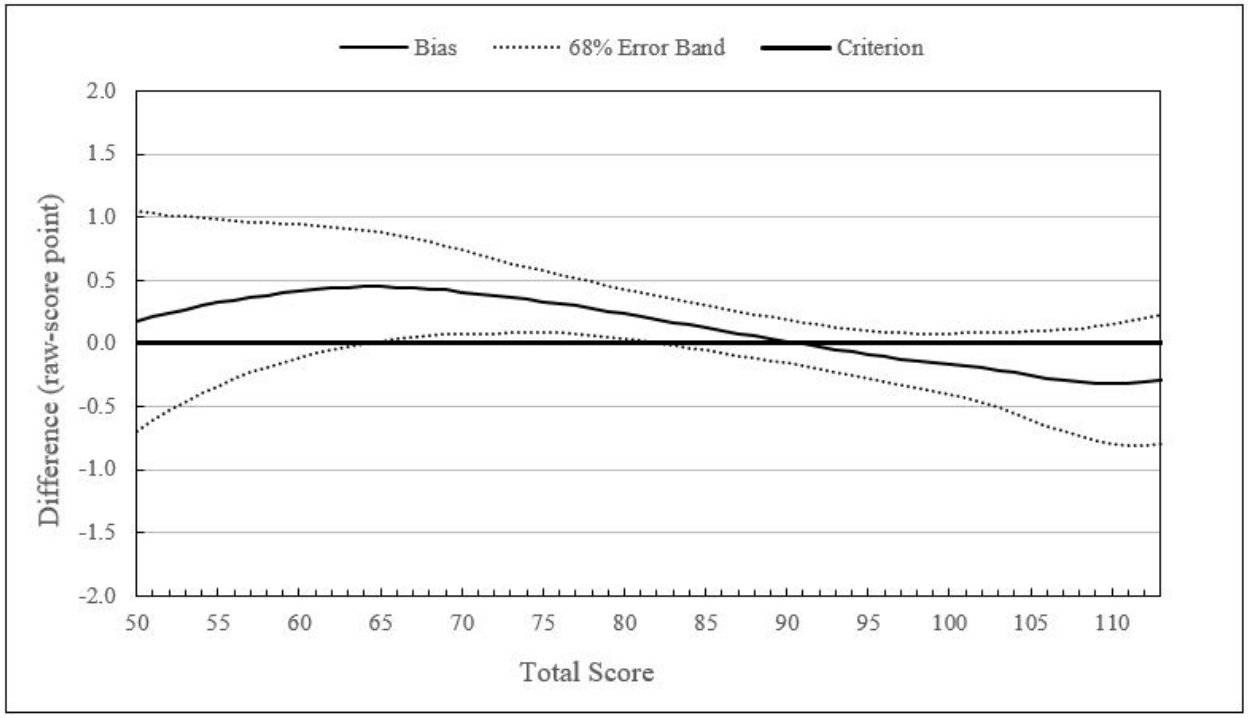

Figure A7. Form A: multiple-choice/constructed-response anchor, rescoring by first scorer (Equating 4).

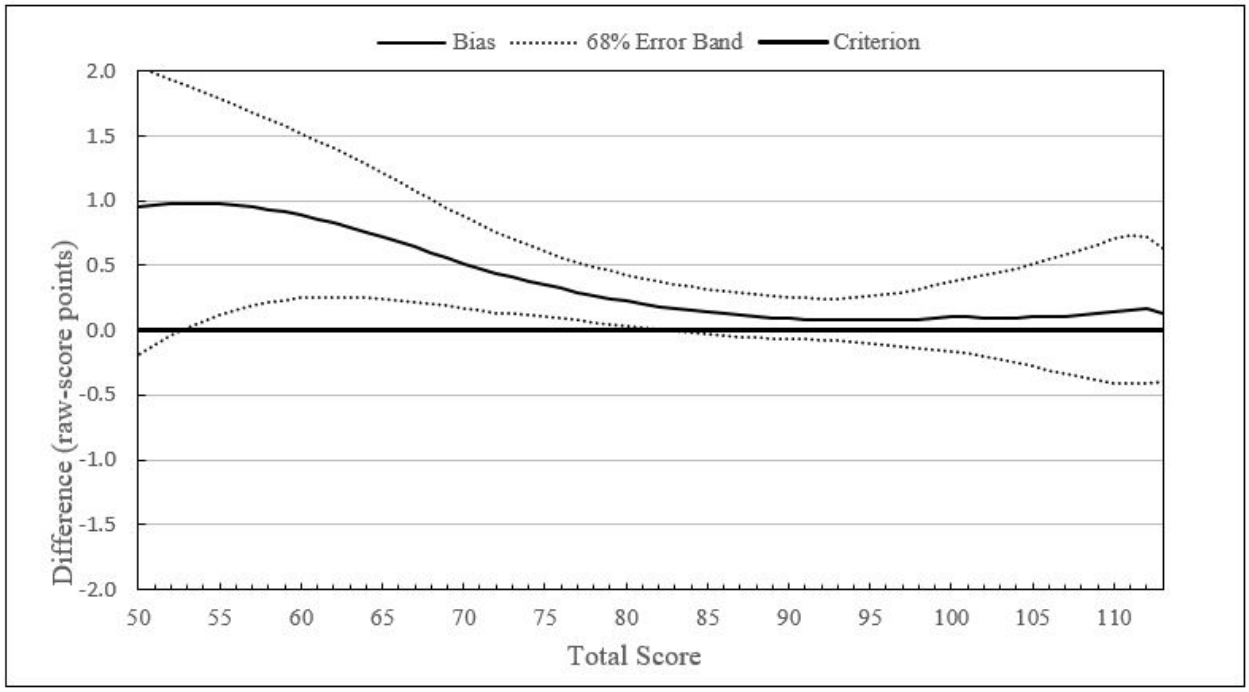

Figure A8. Form A: multiple-choice/constructed-response anchor, rescoring by second scorer (Equating 5).

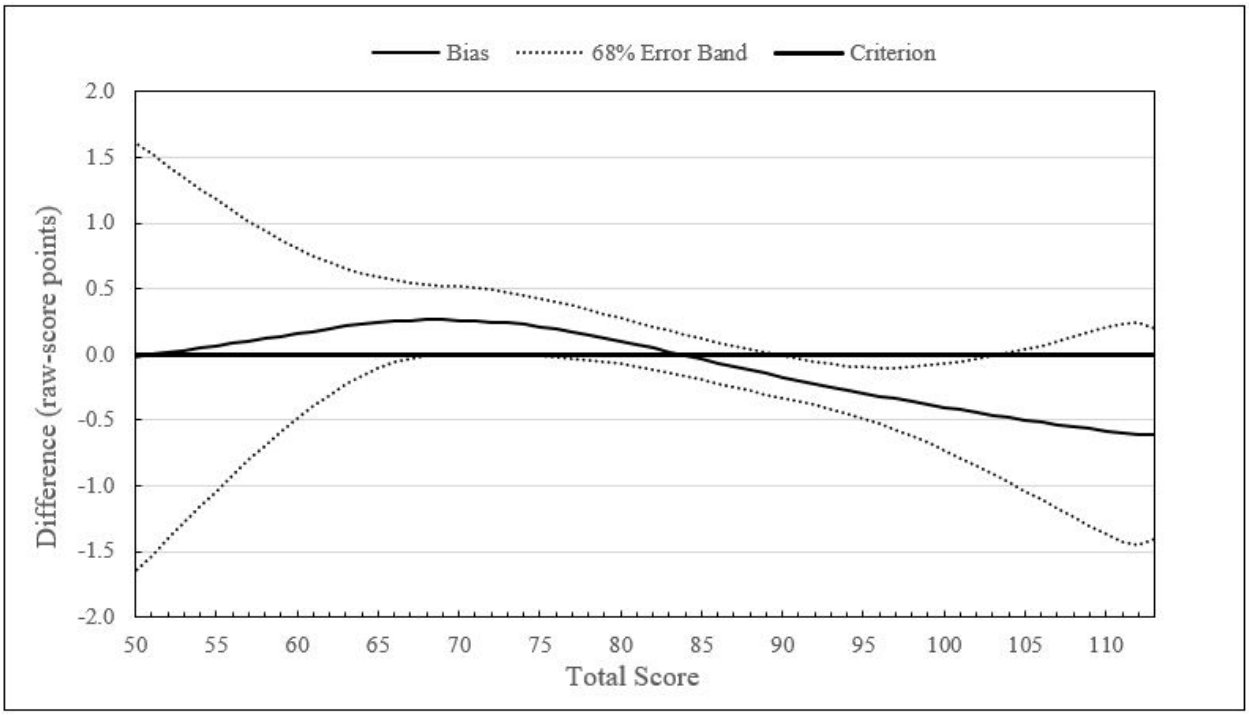

Figure A9. Form B: multiple-choice/constructed-response anchor, rescoring by first scorer (Equating 4).

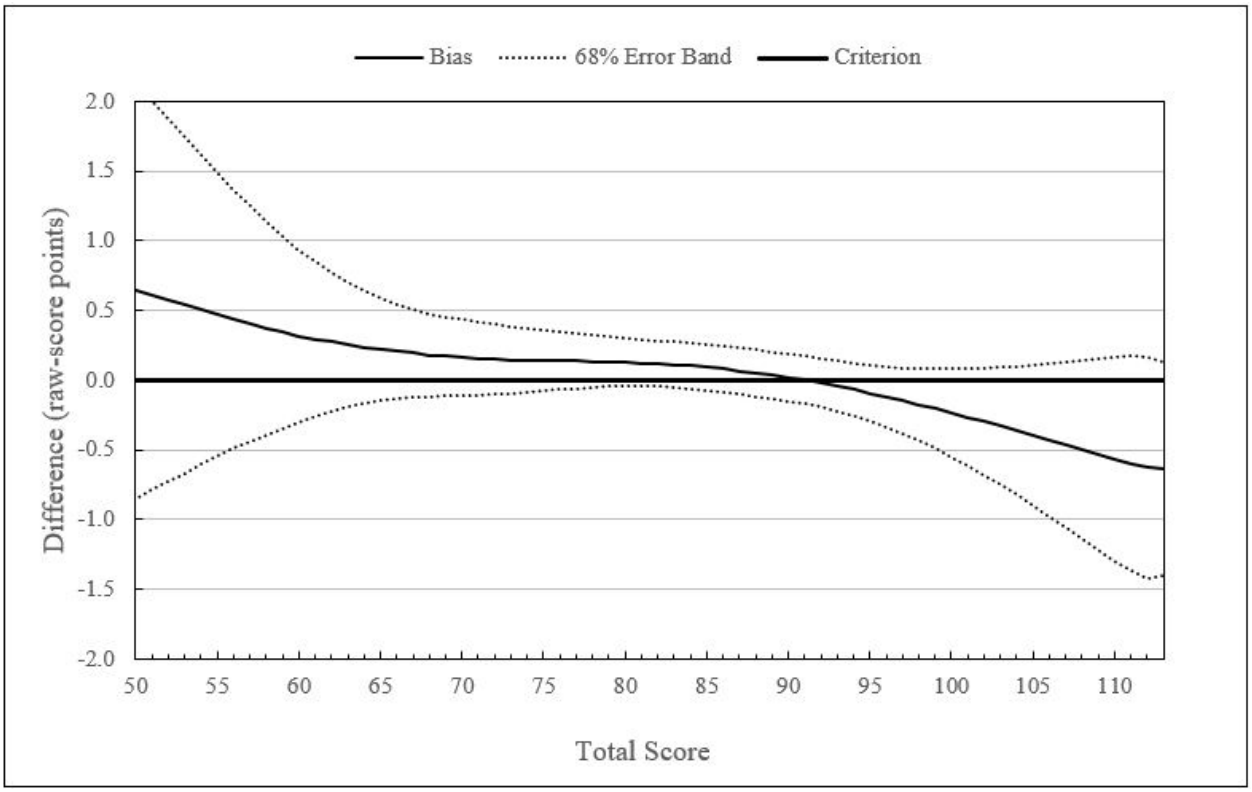

Figure A10. Form B: multiple-choice/constructed-response anchor, rescoring by second scorer (Equating 5).