Research Report

Comparisons Among Designs for Equating Constructed-Response Tests

Sooyeon Kim

Michael E. Walker

Frederick McHale

October 2008

ETS RR-08-53

Comparisons Among Designs for Equating Constructed-Response Tests

Sooyeon Kim, Michael E. Walker, and Frederick McHale
ETS, Princeton, NJ

October 2008

Abstract

This study examined variations of a nonequivalent groups equating design used with constructed-response (CR) tests to determine which design was most effective in producing equivalent scores across the two tests to be equated. Using data from a large-scale exam, the study investigated the use of anchor CR item rescoring in the context of classical equating methods. Four linking designs were examined: (a) an anchor set containing common CR items, (b) an anchor set incorporating common CR items rescored, (c) an external multiple-choice (MC) anchor test, and (d) an equivalent groups design incorporating CR items rescored (no anchor test). The use of CR items without rescoring or the use of an external MC anchor resulted in much larger bias than the other two designs. The use of a rescored CR anchor and the equivalent groups design led to similar levels of equating error.

Key words: Constructed-response test, scoring shift, rater effect, trend scoring method, equivalent groups design, equating

Acknowledgments

The authors would like to thank Robert Smith, Mary Grant, and Daniel Eignor for their helpful comments, and Kim Fryer for the editorial help.

Introduction

For many reasons, large-scale testing programs increasingly use constructed-response (CR) items in their assessments. These items are less susceptible to guessing, they measure production or recall rather than recognition, and they are closer to the types of everyday tasks encountered by examinees (e.g., in the classroom). Along with their benefits, CR items bring certain complications that must be addressed to assure quality and equity in assessments. For example, the group of judges scoring at one time or another can be more or less lenient, and thus CR items can contain more scoring error than multiple-choice (MC) items. Given the widespread use of CR item tests, it is necessary to determine which equating designs will adjust adequately for differences in difficulty and scoring leniency across forms so that scores are fair and accurate. This study examined several procedures for equating CR tests in an attempt to find the most effective procedure.

The Need for Equating

For security reasons, testing programs use multiple forms of the same test in different administrations. In developing various forms of tests, developers use test specifications to ensure that the alternate forms are similar in content and statistical characteristics. For tests containing CR items, the specifications must also include a scoring rubric for each item, which must be consistently applied by the raters when the same items are employed in different test forms or administrations. As well specified as the test development process may be, differences often occur in the statistical difficulty of the alternate forms. We adjust for these differences through the process of equating.

Various equating designs and methods have been discussed thoroughly in the literature (Kolen & Brennan, 2004). A nonequivalent groups with anchor test (NEAT) design has been commonly used to adjust for differences in difficulty among forms that are built to the same specifications. In using a NEAT design, a major drawback with CR tests is the difficulty of identifying a satisfactory anchor test. In many cases, for example, CR items are not reused across different test forms because of ease of memorization (Muraki, Hombo, & Lee, 2000), so no common CR items are available for equating. Some practitioners have suggested using MC items as anchors to adjust for differences in difficulty among test forms containing CR items (e.g., Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). However, evidence suggests that using an all-MC anchor with tests made up of CR items will lead to biased equating results (Kim

& Kolen, 2006; Kim, Walker, & McHale, 2007; Li, Lissitz, & Yang, 1999), possibly because the MC and CR items may measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002).

Even if CR items were reused, and an anchor containing CR items could be found that was representative of and highly correlated with the total test, the anchor may not behave in the same way in both testing groups over time. Scorers could change their scoring standards from one time to the next, so that the anchor items would no longer be equivalent across time. In this situation, the common CR items would not really be equivalent, because different rating teams scored them. Accordingly, applying standard equating practices would lead to erroneous results (Tate, 1999).

A practitioner might see as another possibility equating the tests using a randomly equivalent groups (EG) design. In this situation, no anchor test would be needed. This procedure would be based on the assumption, however, that the previously administered test form, to which the new test form would be equated, behaved identically in the current administration as in the administration in which it, itself, was equated. As mentioned previously, however, changes in scoring severity for the CR items would make this assumption untenable. Recently, Kim, Walker, and McHale (2007) proposed an EG design incorporating rescored CR items to adjust for scoring changes when equating tests are made up of MC and CR items.

Trend Scoring Method

CR items may be difficult to score objectively and reliably. Changes in scoring standards for common CR items are one serious problem that many psychometricians confront in operational situations. Tate (1999, 2000) articulated a solution to the problem of scoring standards that have changed over time in the context of the NEAT design. He suggested a preliminary linking study in which any across-year changes in rater severity could be isolated, so across-group ability differences resulting from the NEAT design could be accurately adjusted and the tests could be properly equated, thereby dealing with anchor from difficulty difference. The linking study involved rescoring responses to the CR anchor items obtained from the old (reference) population. These responses, obtained from the old group of examinees, were rescored by the same raters who scored the responses for the same items for the new group of examinees. Thus, these *trend papers* had two sets of scores associated with them: one from the old set of raters and one from the new rater set. In short, this trend-scoring method entails rescoring responses from the same examinees in order to be able to assess and control for rater severity differences across scoring sessions.

Tate (2003) and Kamata and Tate (2005) used simulation studies to show the effectiveness of the proposed linking method incorporating trend scored papers by employing item response theory (IRT). In practice, however, IRT methods often may not be desirable or advisable in cases of insufficient sample sizes or untenable item-level assumptions. In such cases, classical linear (e.g., chained linear, Tucker, and Levine) and/or nonlinear (e.g., frequency estimation, chained equipercentile) methods would be used to link tests with CR items. Recently, Kim et al. (2007) examined the effectiveness of equating designs incorporating trend scoring using non-IRT linear equating methods with actual data from an operational test. They examined variations of the NEAT design for mixed-format tests, tests containing both MC and CR items, to determine which was most effective in producing equivalent scores across the two tests to be equated. Four equating designs were examined in the context of classical linear equating methods. They used (a) an anchor with only internal MC items, (b) a mixed-format anchor test containing both MC and CR items, (c) a mixed-format anchor test incorporating CR item rescoring, or (d) an EG design with trend scoring, thereby avoiding the need for an anchor test. In their study, the use of the mixed anchor with the NEAT design proved harmful when no trend CR items were incorporated as an anchor in the presence of a change in CR scoring standards. Equating bias caused by a CR scoring shift was controlled, however, through the use of a trend-scoring method.

The Purpose

The study was motivated by the fact that test forms containing CR items will differ in difficulty because either (a) the items are different, or (b) the items are the same but the rater standards have changed. In either of these cases, the test forms should be equated. When both conditions occur, the item and rater effects are completely confounded. Tate (1999, 2000) argued that CR items should be thought of as a function of item/rater combinations. In other words, the same CR item scored using different rater standards should be considered a completely new item. Experience has shown that rater standards will often shift from one administration to another, even if the scoring rubric has not changed and even if the same raters score the same physical items across both administrations. It is important, therefore, to take into account these changes whenever CR items are used as common items.

The present study examined four designs for equating CR item tests with no MC items (i.e., all items are CR) in an attempt to find the most appropriate design. This work is an extension of the previous work conducted by Kim et al. (2007). As mentioned earlier, a major drawback with CR tests is the difficulty of identifying a satisfactory anchor test in a NEAT design. The four equating designs used different anchor sets: (a) a no-trend CR anchor, (b) a trend CR (i.e., rescored) anchor, (c) an external MC anchor, and (d) an EG design with trend scoring, no anchor test required. The purpose of this study was to determine which equating design or designs are most effective in adjusting CR tests for difficulty differences that might be caused by the use of different CR items, changes in the scoring standards for the CR common items, or both.

The present study focused on classical linear equating methods (e.g., chained linear, Tucker, and Levine). Two procedures did not use trend scoring: The first used only external MC items in the anchor; the second used no-trend CR items in the anchor. Two procedures incorporated trend scoring: One used essentially the procedure suggested by Tate (1999), adapted for non-IRT equating methods; the other used an EG design that obviated the need to search for a representative anchor test. The research attempted to answer two major questions: (a) which equating design is the most effective for linking tests with CR items and (b) which anchor test (MC or CR items) works best.

Method

Rating Data

The data for the study were taken from a subject test of a large scale testing program (called Form Z). This test comprised 24 MC and 12 CR items. Four hundred and seventeen examinees and their 12 CR items, scored by Rater Group A, were taken from the reference administration. The same 12 CR items for these 417 examinees were also scored by another set of raters (called Rater Group B) who, in turn, also scored the 12 CR items for a new group of examinees (N = 3,126). Therefore, two independent sets of scores for all CR items were available for those 417 reference examinees, but only a single set of CR scores was available for the 3,126 new examinees. The reference form group (N = 417) only consisted of first-time test takers, but the new form group (N = 3,126) consisted of both repeaters and first-time test takers. Both Rater Group A and Rater Group B were trained by chief readers to score the common CR items using the same procedures and scoring rubrics. The two administrations were about 8 months apart.

Simulated Forms

The data set used in the study was the same used by Kim et al. (2007) in their investigation of equating mixed-format tests. The original test form (Form Z), from which the CR forms in the study were created, has a possible score range of 0 to 72. The score range was derived from 24 MC items, which all total were worth 24 points, and 12 CR items, the rating for which was an integer from 0 to 2 for each item, with all the CR ratings weighted by 2. Two CR forms parallel in both content and difficulty (designated simulated new form and simulated reference form) were created from the original test (Form Z). Figure 1 shows the basic layout for the two parallel forms, along with an external MC test. As shown, the new and reference forms consist of 8 CR items each, with 4 CR items in common, to be used as the anchor in a NEAT design. The possible score ranges for the CR test and anchor were 0 to 32 and 0 to 16, respectively. The 24 MC items were employed as an external MC test in one of the equating designs studied.

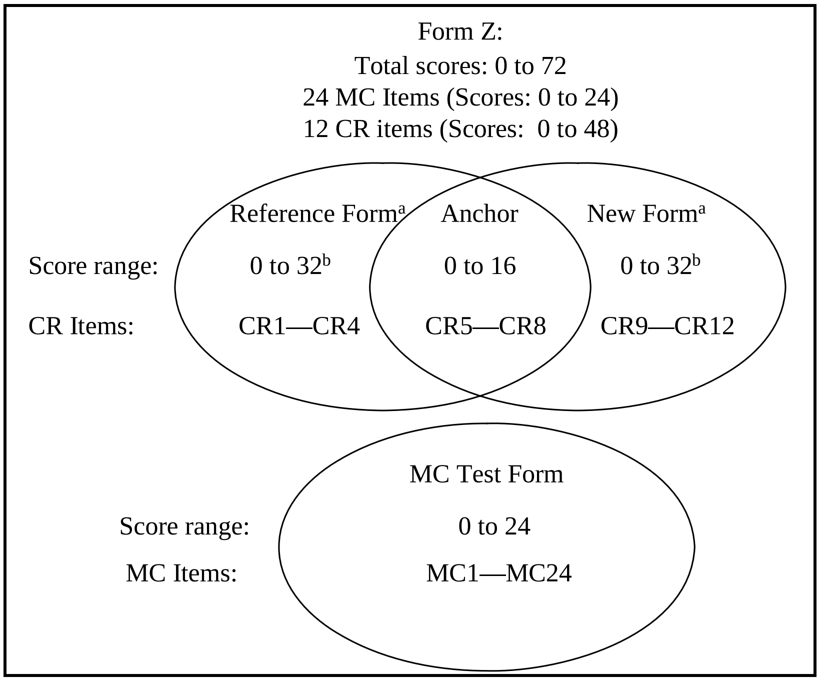


***Figure 1.* Design of the two simulated test forms used in the study.**

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

[a] The CR items in the two forms are actually interspersed throughout the forms and not set in blocks. [b] Each rating for CR items is an integer from 0 to 2 and all the CR ratings are weighted by 2 as in the actual operational situation.

Procedure

Criterion. The study examined ways to place the simulated new form/Rater Group B on scale with the simulated reference form/Rater Group A using different designs and anchor compositions. In the general scenario, the reference form is scored by one set of raters (Rater Group A); whereas the new form, given at a different administration, will be scored by a different set of raters (Rater Group B). We cannot consider the test form separate from the raters who scored the CR items on that form. Thus, if we equated the new form scored by Rater Group A to the reference form scored by Rater Group A, we would still not have the proper scale for the new form that was actually administered: namely, the form scored by Rater Group B. This distinction lies at the heart of the problem this study addresses. Failure to make this distinction can lead to erroneous results. It is essential that the criterion reflect the correct equating of the new test form as given (i.e., scored by Rater Group B) to the reference form as given (i.e., scored by Rater Group A).

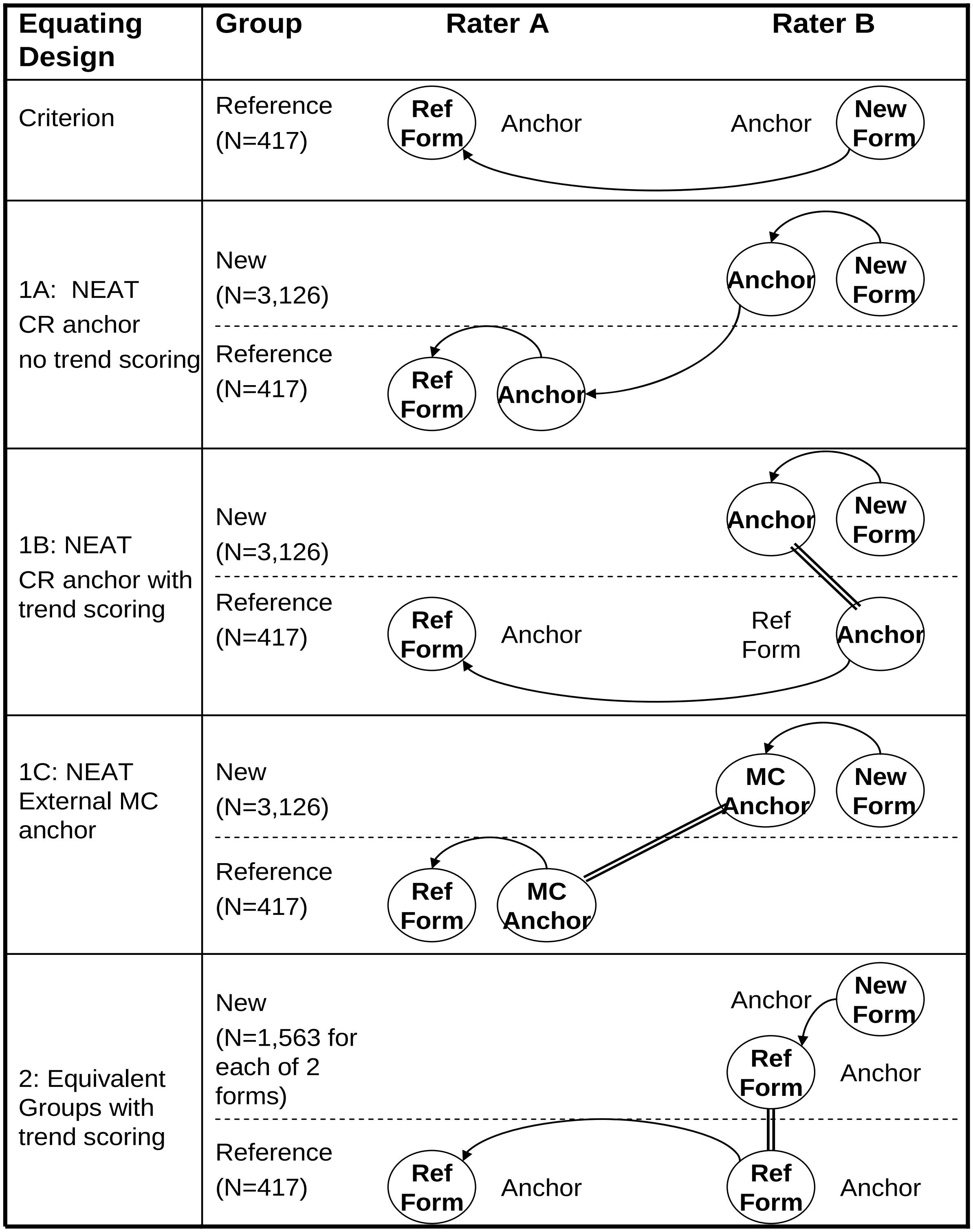
For the 417 reference group examinees, two independent sets of scores for all CR items (new and reference forms) scored by Rater Group A and by Rater Group B were available. Accordingly, the criterion linking was estimated with those 417 examinees using a single group (SG) design. The schematic of this design is presented in the upper section of Figure 2. To estimate the criterion function, total scores on new form/Rater Group B (32 score points) were equated to total scores on reference form/Rater Group A (32 score points) by setting means and standard deviations equal. A smoothed equipercentile equating was also conducted to verify the linearity of the equating relationship. The data were presmoothed using loglinear methods.

Equating designs. Two equating designs were considered in this study. One design is a NEAT design having different types of anchors: (a) a CR anchor and (b) an external MC-only anchor. The other design followed an EG design incorporating CR trend scoring. In both designs, the 417 examinees scored by Rater Group A were the reference form group and the 3,126 examinees scored by Rater Group B were the new form group.

The first design, the NEAT design, is the most common in practice. As shown in Figure 2, in the NEAT design, three different anchor compositions were examined: (a) Design 1A, no trend CR items; (b) Design 1B, trend CR items; and (c) Design 1C, external MC items. In the Design 1A case, the four common CR items were scored by different sets of raters, by Rater Group B in the new form and by Rater Group A in the reference form. Because the trend scoring information was not utilized to adjust for any scoring shift, the success of equating rested on the assumption that Rater Groups A and B used the same scoring standards and applied them consistently. In this case, the common CR scores represented internal anchors in both groups. In Design 1B, however, the four common CR items were scored by the same raters (Rater Group B) in both the reference and the new form groups. Because the CR anchor items for the 417 reference form examinees were also scored by Rater Group B together with the 3,126 new form examinees using a trend scoring method, any CR scoring shift caused by different sets of raters in the reference and new form groups could be adjusted. In the reference form group, the CR anchor scores were external. In Design 1C, the new form was equated to the reference form via an external MC item test.

The second design, an EG design with trend scoring, represented an alternative to the NEAT design. As shown in Figure 2, this design was a combination of an SG design (reference form group) and an EG design (new form group). This design would be possible if a reference form is spiraled with a new form when given to the new form group. The new form group should be randomly split among the new form and the reference form to obtain an EG design. Because the trend scoring method is still applied in this design, old papers randomly selected from the reference form group should be rescored by the same set of raters who score the new form sample. In this design, the inclusion of an anchor test in the two spiraled forms is not necessary, although the use of a common block of items across the two spiraled forms could enhance the accuracy of the equating function. In this study, the two spiraled forms have 4 CR items in common, by design.

To carry out the second design, the reference form was rescored by Rater Group B via a trend scoring procedure (i.e., by inserting papers of the 417 examinees into the rating process for the 3,126 new form examinees). An adjustment for rater severity is then made by linking the reference form scored by Rater Group B to the same reference form scored by Rater Group A using an SG design (*N* = 417), which adjusts for trend. The second component of this design was an EG design where the reference form and new form were randomly assigned in the new form group of 3,126 examinees (new form [*N* = 1,563], reference form [*N* = 1,563]). The CRs in the new and reference forms are all scored by Rater Group B. The total score on the new form was then equated to the reference form using an EG design.



Equating Design	Group	Rater A	Rater B
Criterion	Reference (N=417)	Ref Form Anchor	Anchor New Form
1A: NEAT CR anchor no trend scoring	New (N=3,126)		Anchor New Form
	Reference (N=417)	Ref Form Anchor	
1B: NEAT CR anchor with trend scoring	New (N=3,126)		Anchor New Form
	Reference (N=417)	Ref Form Anchor	Ref Form Anchor
1C: NEAT External MC anchor	New (N=3,126)		MC Anchor New Form
	Reference (N=417)	Ref Form MC Anchor	
2: Equivalent Groups with trend scoring	New (N=1,563 for each of 2 forms)		Anchor New Form Ref Form Anchor
	Reference (N=417)	Ref Form Anchor	Ref Form Anchor

***Figure 2.* Schematic of the criterion and equating designs examined in this study.**

Note. CR = constructed response, MC = multiple-choice, NEAT = nonequivalent groups with anchor test, Ref = reference.

Evaluation

As mentioned previously, for all equating designs, linear equating methods (e.g., chained linear, Tucker, and Levine) were used. Many observed-score equating methods employ a linear equating function. All these functions and their (untestable) assumptions are described in detail elsewhere (Kolen & Brennan, 2004; Livingston, 2004; von Davier & Kong, 2005). The new form equated raw score conversion obtained using each equating method in each equating design were compared with the criterion conversion. The differences among the conversions were quantified using the root mean squared difference (RMSD),

$$RMSD = \sqrt{\sum_{i=0}^{32} w_i \left[\hat{e}_i\left(x_i\right) - e_i\left(x_i\right)\right]^2}, \tag{1}$$

where *i* represents a raw score point, $\hat{e}_i\left(x_i\right)$ is the equated scores of an equating method in a design at raw score *x*, $e_i\left(x_i\right)$ is the criterion equating function at raw score *x*, and w_i is the relative proportion of the new form examinees at each score point.

Furthermore, standard errors of equating (SEE) and estimates of bias were generated using a resampling technique. A total of 500 bootstrap samples (i.e., 500 replications) were obtained in each equating design using the SAS PROC SURVEYSELECT procedure that randomly selects units *with replacement*. In each replication, examinees were randomly drawn *with replacement* from each reference and new form group until bootstrap samples consisted of exactly the same number of examinees as in the actual reference (*N* = 417) and new (*N* = 3,126 in the NEAT design; *N* = 1,523 in the EG design) form groups. Then the new form scores were equated to the reference form for those 500 samples in each equating design. In this case, equating bias was defined as the mean difference between an equating method and the criterion equating over 500 replications. The standard deviation of these differences at each score point over 500 replications was used as a measure of the conditional standard error of equating (CSEE) or error due to sampling variability. The sum of squared bias and squared CSEE was considered an indication of total undesirable equating variance at each score point, and the square root of this value defined the conditional root mean squared error (RMSE) index. The following equations represent bias, equating error (CSEE), and RMSE measures conditioned on each raw score point (x_i):

$$Bias_i = \overline{d}_i = \frac{\sum_{j=1}^{J}\left[\hat{e}_j\left(x_i\right) - e\left(x_i\right)\right]}{J}, \tag{2}$$

$$CSEE_i = s(d_i) = \sqrt{Var_j\left[\hat{e}_j\left(x_i\right) - e\left(x_i\right)\right]} = \sqrt{Var_j\left[\hat{e}_j\left(x_i\right)\right]}, \tag{3}$$

$$RMSE_i = \sqrt{\overline{d}_i^{\,2} + s(d_i)^2}, \tag{4}$$

where *j* is a replication, *J* is the total number of replications (500), $\hat{e}_j(x)$ denotes the raw score equivalent calculated from an equating function (design) in the sample *j*, and d_i is the difference between $\hat{e}_j\left(x_i\right)$ and $e\left(x_i\right)$.

As overall summary measures, we computed the weighted average root mean squared bias, $\sqrt{\sum_i w_i Bias_i^2}$; the weighted average standard error of equating, $\sqrt{\sum_i w_i CSEE_i^2}$; and the weighted average RMSE, $\sqrt{\sum_i w_i RMSE_i^2}$, across the new form group score distribution, where w_i is the relative proportion of the new form examinees at each score point.

Results

Criterion

Total test scores on the new form were equated to total test scores on the reference form with a total of 417 examinees based on an SG design to define the criterion. The CR scores on the new and reference forms were generated by Rater Groups A and B, respectively. The means and standard deviation were 20.86 and 4.74 for the reference form and 20.99 and 4.68 for the new form, respectively. As the first step, for each raw score on the new form, the equivalent raw score on the reference form was determined using the mean-sigma (linear) and direct equipercentile (nonlinear) methods. Figure 3 presents equated raw score differences between the mean-sigma and direct equipercentile methods. The differences between the two functions appeared to be substantial for the lower end of raw scores, but almost no data were available in that region of the score scale. Because the differences between two equating functions were considered negligible where most of the examinees were located, the linear function was used as the criterion and was compared with the equating functions derived from various equating designs for our research purposes. This decision seemed to be reasonable in that the criterion functions were derived from a relatively small sample (*N* = 417).

Anchor Design

Summary statistics of total and anchor scores for examinee groups taking new and reference forms are presented in Table 1. The total score mean of the reference form group (*M* = 20.86) was higher than that of the new form group (*M* = 19.27). Regarding the anchor, the reference form group showed higher means for both anchor formats than did the new form group, because the reference form group consisted of first-time test takers who tended to score higher than test repeaters.

In the external MC anchor test design, the magnitude of the correlations between the total CR test score and external MC anchor scores was relatively low (*r* = .46 –.47) but fairly similar in both groups. The magnitude of the correlations between the total CR test score and no-trend CR anchor scores was high (*r* = .82–.84) and very similar in both groups. In the trend CR anchor design, however, the CR anchor was correlated more highly with the total CR score in the new form group (*r* = .84) than in the reference form group (*r* = .57); here the CR anchor was internal for the new form group but external for the reference form group. As explained previously, for the trend scoring method, the CR anchor scores were generated at a different time by a different set of raters (Rater Group B), and accordingly, the CR anchor scores were not part of the total CR scores.

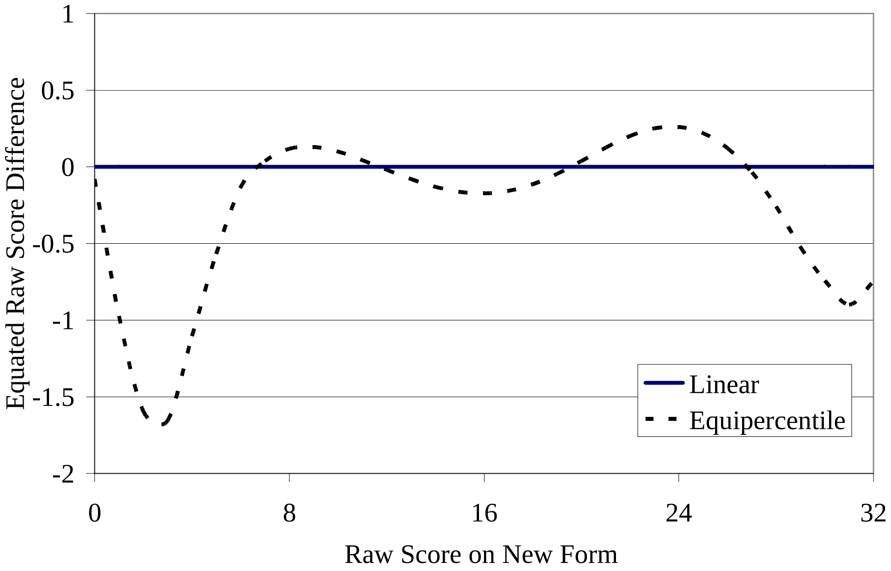


Figure 3. **Difference plot between linear and direct equipercentile criterion functions.**

Table 1

Summary Statistics for Examinee Groups Taking New and Reference Forms in the Anchor Design

	Equating designs					
	Design 1A: No-trend CR anchor		Design 1B: Trend CR anchor		Design 1C: External MC anchor	
	Reference group	New group	Reference group	New group	Reference group	New group
Number of examinees	417	3,126	417	3,126	417	3,126
Number of CR items	8	8	8	8	8	8
CR rater group	Rater A	Rater B	Rater A	Rater B	Rater A	Rater B
Total score mean	20.86	19.27	20.86	19.27	20.86	19.27
Total score standard deviation	4.74	5.46	4.74	5.46	4.74	5.46
Number of anchor items	4	4	4	4	24	24
CR anchor rater group	Rater A	Rater B	Rater B	Rater B	NA	NA
Anchor score mean	10.19	9.89	10.82	9.89	19.16	18.31
Anchor score standard deviation	3.01	3.21	2.79	3.21	3.02	3.54
Correlation of total and anchor scores	.82	.84	.57	.84	.47	.46

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

Table 2 presents the differences between each linear equating function and the criterion, using the RMSD deviance measure. Looking at Table 2, among the three linear methods, the Levine method yielded the smallest RMSD and the Tucker method yielded the largest RMSD, regardless of anchor type. To the extent that the anchor-total correlations depart from 1.00, Tucker equating adjusted as if the equating samples were more similar in ability than the anchor scores indicated. As a result, we expected Tucker equating to be biased, because the reference and new form groups differed substantially in this study. For the anchor test design, the use of trend CR items in the anchors greatly improved equating. For all three linear methods, RMSD values were much smaller in the trend CR anchor case than in both the external MC and no-trend CR anchor cases. Incorporating no-trend CR anchor information into the estimation of equating functions seems to be problematic unless CR scoring standards are well maintained over time by human raters.

Figure 4 plots the conditional equated score difference between the chained linear equating function and the criterion in each equating design. For simplicity sake, only the results from the chained linear equating function are presented in Figure 4. Plots looking at the same differences for Tucker and Levine were similar in nature. The no-trend CR anchor case showed the largest difference over almost all the raw scores. Again, this result clearly indicated the potential problems caused by using a no-trend CR anchor in a NEAT design. The difference between the trend CR anchor and external MC anchor was negligible for the raw score range of 0 to 16, but the trend CR anchor case showed smaller differences than the external MC case for the raw score range of 16 to 32, where most the examinees were located.

Table 3 presents the summary of the weighted average root mean squared bias, equating error, and RMSE derived from a bootstrap resampling technique for each equating design. We examine here the differences between the linear criterion and the chained linear equating results with regard to equating bias, error, and RMSE indices. For simplicity, only the chained linear method was employed. As shown, the use of a no-trend CR anchor resulted in the smallest error but the largest bias, leading to the largest RMSE. The use of external MC anchor yielded a much larger bias and also a slightly larger equating error than the trend CR anchor case did. The trend CR anchor yielded the smallest bias, leading to the smallest RMSE, compared to the other two anchor design cases. Although equating error was fairly comparable for the three designs, the magnitude of bias was substantially larger in both the external MC and no-trend CR anchor cases than in the trend CR case.

Table 2

Summary of Root Mean Squared Difference (RMSD) Between Three Models of Linear Equating Results and the Criterion for Each Equating Design

	Equating method		
Equating design	Chained linear	Tucker	Levine
NEAT:			
Design 1A: No-trend CR anchor	1.357	1.471	1.214
Design 1B: Trend CR anchor	0.181	0.876	0.180
Design 1C: External MC anchor	0.415	1.299	0.366
Equivalent groups with trend scoring	0.161	--	--

Note. The Tucker and Levine methods, which incorporate the correlation information between the two forms into the estimation of the equating function, were not applicable in the equivalent groups with trend scoring case. For that reason, only the result from the chained linear method was examined for this design. CR = constructed response, MC = multiple choice, NEAT = nonequivalent groups with anchor test.

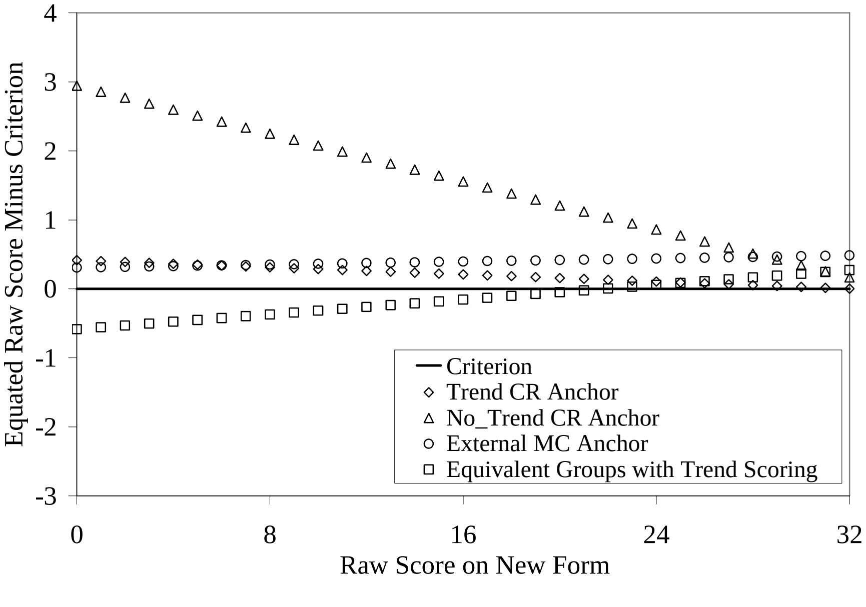


Figure 4. **Difference between chained linear equating and the criterion in the four equating designs.**

Table 3

Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for Each Equating Design, With Chained Linear Equating

	Deviance measure		
Equating design	Bias	Equating error	RMSE
NEAT:			
Design 1A: No-trend CR anchor	1.352	0.206	1.367
Design 1B: Trend CR anchor	0.180	0.321	0.368
Design 1C: External MC anchor	0.446	0.403	0.601
Equivalent groups with trend scoring	0.231	0.346	0.416

Note. CR = constructed response, MC = multiple choice, NEAT = nonequivalent groups with anchor test.

Figures 5 to 7 plot the conditional bias of chained linear equating in the anchor design, along with an error band representing plus or minus one empirical CSEE (i.e., 68% error band). The error band for chained linear equating was slightly wider in the external MC anchor case than in the other CR anchor cases, particularly for the raw score range of 0 to 10. When some CR items are common across test forms, the results indicate that the trend CR anchor would provide better results than the external MC anchor.

Equivalent Groups Design With Trend Scoring

Summary statistics of the total scores for examinee groups taking the reference and new forms in the EG design are presented in Table 4. In this design, the new form sample (*N* = 3,126) was randomly split to simulate the spiraling of the new form with the reference form. Because the new and reference forms were nearly parallel, the total means were similar for both groups. For the 417 examinees who took the reference form, two sets of total scores were available because two different sets of raters, Rater Groups A and B, scored their CR items at different times. The reference form scores for the reference sample (*N* = 417) and for the new form sample (*N* = 1,563) could be directly compared because the same Rater Group B generated scores for the CR items in both cases. Again, the examinee group of 417, composed of only first-time test takers, was more proficient than the two new form groups, which included test-repeaters as well.

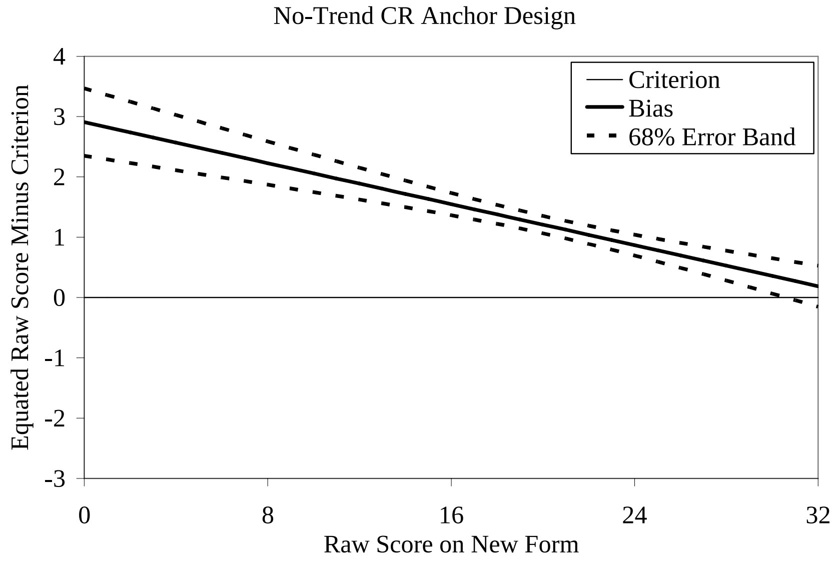


Figure 5. **Difference of chained linear equating from the criterion, for the nonequivalent groups anchor test design with no-trend CR items in the anchor.**

Note. Solid horizontal line at zero is the criterion. Dashed lines show the bootstrapped standard error of equating.

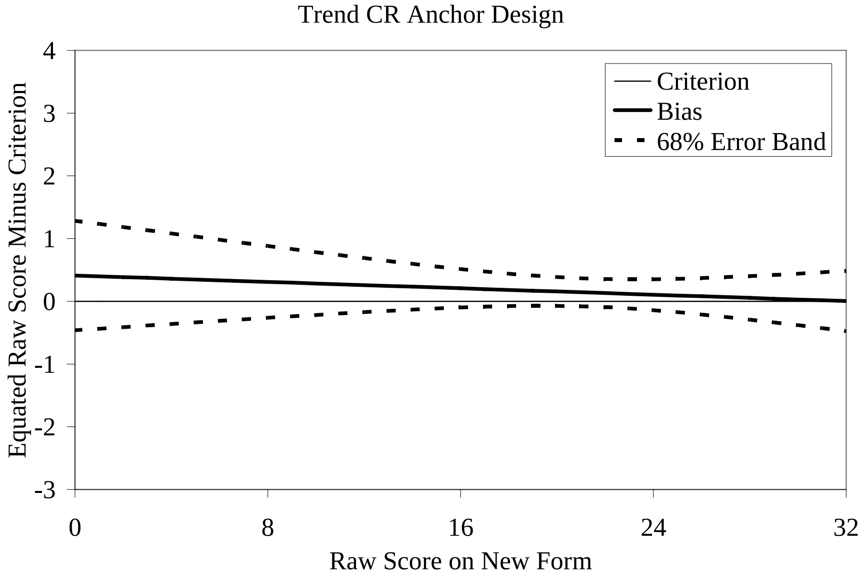


Figure 6. **Difference of chained linear equating from the criterion, for the nonequivalent groups anchor test design with trend constructed-response (CR) items in the anchor.**

Note. Solid horizontal line at zero is the criterion. Dashed lines show the bootstrapped standard error of equating.

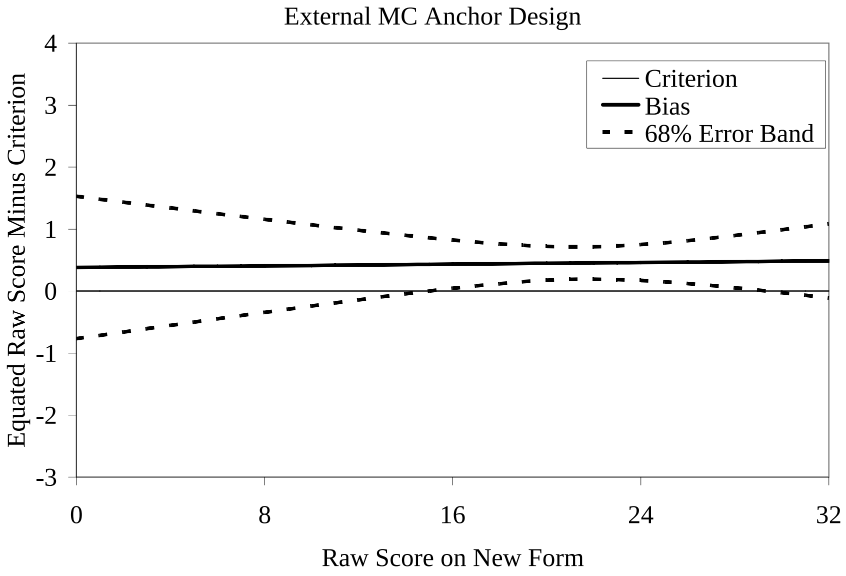


***Figure 7.* Difference of chained linear equating from the criterion, for the nonequivalent groups anchor test design with external multiple-choice (MC) items in the anchor.**

Note. Solid horizontal line at zero is the criterion. Dashed lines show the bootstrapped standard error of equating.

Table 4

Summary Statistics for Examinee Groups Taking Reference and New Forms in the Equivalent Group Design With Trend Scoring

	Reference group	Reference group	New group	New group
	Reference form	Reference form	Reference form	New form
	CR rater group			
	A	B	B	B
Sample size	417	417	1,563	1,563
Number of CR items	8	8	8	8
Mean	20.86	21.50	19.58	19.19
Standard deviation	4.74	4.79	5.70	5.42

Note. CR = constructed response.

The summary statistics for the EG with trend scoring design are also summarized in Tables 2 and 3. In this design, the first half of the chain linked the total scores on the reference form generated by Rater Group B to the total scores on the reference form generated by Rater Group A with the 417 examinees using an SG design. The second half of the chain linked the total scores on the new form generated by Rater Group B to the total score on the reference form generated by Rater Group B using an EG design. Although there was no anchor test in an EG design, the equation used to obtain the linear equating result was indistinguishable from the equation for the chained linear equating in the NEAT design case. The Tucker and Levine methods, however, which incorporated the correlation information between the two forms into the estimation of the equating function, were not applicable in this case. Accordingly, only the result from the chained linear method was examined for the EG design.

The EG design with trend scoring performed as well as the trend CR anchor NEAT design, leading to similar levels of RMSD, bias, error, and RMSE. The EG design yielded the smallest RMSD of the four designs when the chained linear method was used. The EG design resulted in slightly larger bias and error, leading to a larger RMSE, however, than did the trend CR anchor design. Among the four designs, the trend CR anchor yielded the smallest bias and RMSE values. Figure 8 plots the conditional bias of chained linear equating in the EG design, along with an error band representing plus or minus one empirical CSEE (i.e., 68% error band). The conditional bias was negligible across all the raw score points except the low end of scores.

Conclusions

The present study examined the efficacy of a number of designs for equating tests composed exclusively of CR items using classical equating methods. The different equating designs, incorporated (a) a trend scored CR anchor, (b) a CR anchor that was not trend scored, (c) an external MC anchor, or (d) no anchor at all. The findings of the present study paralleled previous findings for the mixed-format tests (Kim et al., 2007), and thus many practitioners may consider them when creating equating plans for tests that employ CR items.

For tests that use CR items, scoring consistency over time should be investigated to ensure the accuracy of examinees' scores. Many practitioners may overlook the difference in CR scoring standards across test form administrations and attempt to use conventional equating methods, ignoring CR scoring differences. However, the use of the CR anchor might be harmful when no-trend

CR items are incorporated as an anchor in the presence of a change in CR scoring standards. The use of no-trend CR items in the presence of a change in CR scoring standards will result in serious equating bias. In the case of examinations using cut scores, the use of traditional equating methods using no-trend scored CR anchors could result in incorrect pass/fail decisions for examinees.

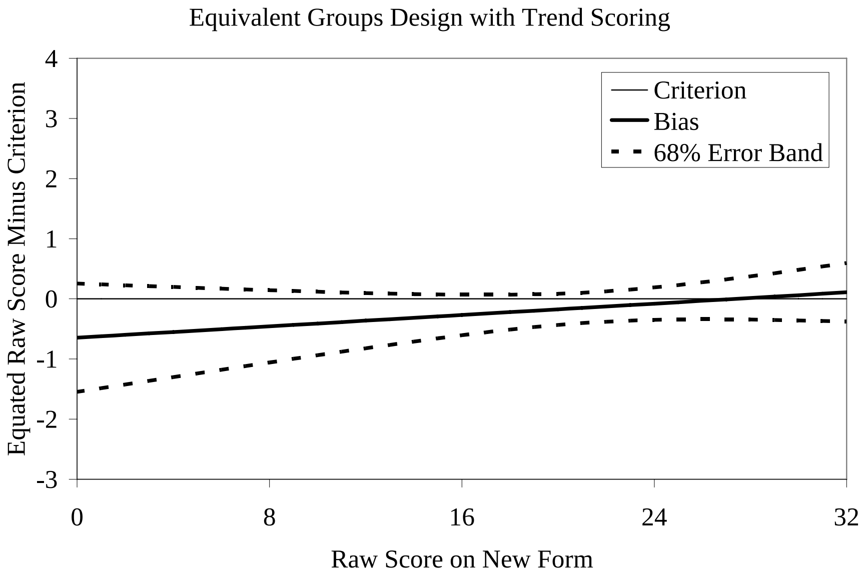


***Figure 8.* Difference of chained linear equating from the criterion, for the equivalent groups design with trend scoring.**

Note. Solid horizontal line at zero is the criterion. Dashed lines show the bootstrapped standard error of equating.

As mentioned previously, some practitioners have suggested using MC items as anchors to control for differences among test forms containing CR items (e.g.,. Baghi et al., 1995; Ercikan et al., 1998). This format of equating may be inappropriate, though, due to the low correlations between the CR and MC scores, indicating that the MC and CR scores measure different constructs (Bennett et al., 1991; Sykes et al., 2002). Previous research showed that use of MC-only anchors with CR tests could result in potentially large equating bias (Li et al., 1999; Kim & Kolen, 2006; Kim et al., 2007). The present study is consistent with those previous findings. The external MC-only anchor design produced quite large RMSD, bias, equating error, and RMSE, showing inferiority to the trend CR anchor and EG designs. The external MC anchor design, which is commonly used in practice, should therefore be used with caution, unless the correlation between CR and MC is substantially high.

As in the previous study that looked at mixed format tests (Kim et al., 2007), this study showed that equating bias caused by a scoring shift could be controlled by using a trend scoring method. The trend scoring method is expensive and difficult to implement in practice; however, in an image or online scoring system with proper tools, its implementation would be straightforward. The trend scoring method has statistical strengths in detecting a CR scoring shift. The trend CR anchor displayed much better performance than did the no-trend CR anchor in recovering the criterion equating function, primarily because of the bias reduction. The equating error was actually slightly larger for the trend CR anchor than that for the no-trend CR anchor. This may be attributable in great part to the somewhat lower anchor-total correlation for the reference form (because the anchor is treated as being external to the test) when trend scoring is used. The slight increase in error was more than offset by the decreased bias, resulting in lower overall RMSE for the trend CR anchor.

In many operational situations, the anchor design with either CR (trend or no-trend) or external MC anchor test is the current practice for the CR-only test. In this study, the trend CR anchor design displayed superior performance to the other two anchor designs and even slightly better performance than did the EG with trend scoring (possibly no anchor) design. The superiority of the EG design observed for the mixed-format test (Kim et al., 2007) was not salient for the CR-only test.

In general, the superiority of the EG design over the NEAT design is that the representativeness of the anchor becomes irrelevant because the anchor is unnecessary in the EG design. The equating sample around 3,000 examinees, however, may not be large enough to spiral two forms under the EG design framework. Although the new and reference forms are close to parallel in this study, the mean of the reference form group was slightly higher than that of the new form group, implying insufficiency of random equivalence across the two groups. Such insufficiency may lead to equating bias, which was the case in this study. In addition to that, the EG design may not be feasible in many testing situations.

Overall, the observed differences in performance between the trend CR anchor design and the EG design is not great. There are tradeoffs between the two designs, however, that may make one design preferable to the other. For example, some items need to be common to both test forms to use the trend CR anchor design, but this requirement is not necessary for the EG design. On the other hand, only common items need to be rescored in the trend CR anchor design, but all CR items should be trend scored in the EG design. Only the new test form needs to be administered in the trend CR anchor design, but both test forms (i.e., new and reference) should be spiraled in each administration if the EG design is used. Finally, in principle an EG design requires a substantially larger number of examinees than a NEAT design to achieve the same level of equating error.

Given the limitations listed above, practitioners may choose one or the other of the NEAT or EG designs, depending upon the situation. The NEAT design may be preferred when the number of CR items to be scored must be kept as small as possible; when sample sizes in each administration are relatively small, such that spiraling would result in insufficient numbers of examinees to ensure random equivalence or sufficiently small equating error; or when security issues or other concerns preclude readministering the entire reference form. The EG design may be preferred in cases in which anchor tests are not feasible or may not be content representative: for example, for set based tests in which the anchor would need to be an intact group of interdependent items. The EG design might also allow equating of short all-CR tests; or other tests in which no CR items are reused.

An issue often overlooked in operational settings occurs when a test is reprinted or reused in an intact fashion. In this case, the original test score conversion, obtained when the test was first equated, is applied in subsequent administrations. While this may work fine for tests composed exclusively of MC items, as shown in the previous study (Kim et al., 2007), the present research demonstrates that it clearly poses problems for tests that contain CR items. In this case, the original test score conversion may no longer apply. The reason is that the scoring standards may not be constant across administrations. Thus, even in the case of test form reuse, trend scoring should be implemented. If the trend scoring indicates that a rater shift has taken place, the CR reprint form should be treated as a new form and re-equated to adjust for differences in rater severity.

There are limitations, however, in generalizing the findings of the current study in practice. This study is based solely on a single test form with a single administration. The criterion was derived from relatively small samples (fewer than 500 examinees). Additional empirical evidence about the trend CR anchor or an EG design should be gathered using various data sets from different formats, different subject tests, and different administrations to enhance its generalizability.

References

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

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

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*(2), 137–154.

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42*(2), 193–213.

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

Kim, S., Walker, M. E., & McHale, F. (2007, April). *Equating of mixed-format tests in large scale assessments.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999, April). *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, Quebec, Canada.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

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

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

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*(4), 336–346.

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

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement*, *63*(6), 893–914.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the Non-equivalent group design. *Journal of Educational and Behavioral Statistics, 30*, 313-342.