Equating of Mixed-Format Tests in Large-Scale Assessments

Sooyeon Kim

Michael E. Walker

Frederick McHale

May 2008

ETS RR-08-26

Equating of Mixed-Format Tests in Large-Scale Assessments

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

May 2008

Abstract

This study examined variations of the nonequivalent-groups equating design for mixed-format tests—tests containing both multiple-choice (MC) and constructed-response (CR) items—to determine which design was most effective in producing equivalent scores across the two tests to be equated. Four linking designs were examined: (a) an anchor with only MC items; (b) a mixed-format anchor containing both MC and CR items; (c) a mixed-format anchor incorporating CR item rescoring; and (d) a hybrid combining single-group and equivalent-groups designs, thereby avoiding the need for an anchor test. Designs using MC items alone or those using a mixed anchor without CR item rescoring resulted in much larger bias than the other two design approaches. The hybrid design yielded the smallest root mean squared error value.

Key words: Mixed-format test, scoring shift, trend-scoring method, hybrid no-anchor design, mixed-format anchor, equating

Acknowledgments

This study was funded by the 2006 Equating, Scaling, Applied Psychometrics Research Initiative. The authors would like to thank Mary Grant, Catherine McClellan, and Marna Golub-Smith for their helpful comments and Kim Fryer for the editorial help.

Table of Contents

Page

List of Tables

List of Figures

Page

Introduction

This research examined several procedures for equating mixed-format tests—tests including both constructed-response (CR) and multiple-choice (MC) items—in an attempt to find the most effective procedure. Such research is necessary because large-scale testing programs are increasingly using CR items in their assessments. By *CR item*, we mean any item in which the examinee must produce a response to the item, rather than just selecting the correct answer from a list of possible options.

In a sense, the inclusion of CR items in tests marks a return to the roots of standardized testing. The first test given by the College Entrance Examination Board (or College Board) in 1901, for example, comprised all-essay items in nine subjects. Not until 1937 did the College Board introduce all multiple-choice tests (Donlon, 1984). Proponents of using CR test items in assessments view them as more authentic than MC items (Haertel & Linn, 1996), because CR items more closely resemble the real-world tasks associated with the construct to be measured; therefore, the use of CR items should lead to minimal construct underrepresentation (Messick, 1994). Furthermore, proponents argue, whereas MC items measure simple recognition of information, CR items measure higher-order thinking processes (Baker, O'Neil, & Linn, 1993).

The distinction between MC and CR items does not involve a dichotomy, but a continuum representing different degrees of structure versus open-endedness in the response (Messick, 1996). CR items can range from grid-in mathematics problems, to items requiring a short written response or essay, to complex performance assessments. In general, along with structured response comes the possibility of more objective scoring; conversely, the price for open-endedness is usually the necessity of basing item scores on some form of subjective judgment. Thus, in large-scale assessments, increases in scoring time, cost, and complexity, along with the possibility of there being inadequate levels of reliability, usually accompany the decision to include CR items.

In summary, MC items are economically practical and ensure objective and reliable scoring. CR items may be difficult to score objectively and reliably, but they may measure examinees' understanding of that particular content at a deeper level than MC items. Because both MC and CR items display strengths as well as weaknesses, it is typical that many assessments tend to be of mixed format, including both MC and CR items.

The Need for Equating

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 scoring rubrics for each item, which must be consistently applied by the raters across different test forms and administrations. As well specified as the test development process may be, differences often occur in the statistical difficulty of the alternate forms (e.g., because the items are slightly easier or harder than expected). 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). These can be readily applied to the equating of tests with MC items. Several item response theory (IRT) linking methods have also been extended for use with tests containing CR items (Baker, 1992; Cohen & Kim, 1998; Kim & Kolen, 2006). Nevertheless, applying these methods to tests that include CR items may still be problematic. For one thing, CR items take longer to answer than MC items, and therefore a test comprised entirely of CR items will necessarily be shorter (and less reliable) than a test with all MC items that is administered in the same length of time. Factors such as test length and reliability directly affect the quality of equating (Fitzpatrick & Yen, 2001).

Perhaps the most commonly used equating design involves the use of nonequivalent groups of examinees with common (anchor) items, called a *NEAT design*, for adjusting test scores. 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 they are easy to memorize (Muraki, Hombo, & Lee, 2000), and thus no common CR items are available for equating. Some practitioners have suggested using MC items as anchors to control for differences among test forms containing CR items (e.g., see Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). However, evidence suggests that using an all-MC anchor will lead to biased equating results (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999); possibly because 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 a CR anchor 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 group to the next, so that the anchor items would no longer be equivalent across groups. In this situation, 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 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 it did 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.

Tate (1999, 2000) articulated a solution to the problem of subjective or changing scoring standards in the context of the nonequivalent-groups anchor test (NEAT) design. He suggested a preliminary linking study in which any across-year changes in rater severity could be isolated, so that across-group ability differences could be accurately assessed and the tests could be properly equated. The linking study involves rescoring responses to the CR anchor items obtained from the old (reference) population. A representative sample of anchor item papers for examinees from Year 1 (reference) is inserted into the rating process for Year 2 (new). These responses, obtained from the old group of examinees, are scored by the same raters scoring responses for the same items for the new group of examinees. Thus, these *trend papers* have two sets of scores associated with them: one from the old set of raters and one from the new rater set.

Tate (1999) explained the year-to-year linking procedure for tests comprised of polytomous items in the context of IRT-based linking. Consider the reference form, which has been placed on scale in Year 1, such that for the MC items the item parameter estimates are based on the Year 1 examinees. For the CR items, however, Tate holds that it is incorrect to think of item parameters. Rather, we must consider item/rating team parameters. The original parameter estimates for the reference form are based on Team 1, who rated the CR responses in Year 1. If the rating team changes, the parameter estimates will be different. Therefore, the rating team must be taken into account when equating.

An intermediate goal in the process of equating a new form given in Year 2 to a reference form given in Year 1 is to express the ability of the Year 2 examinees in the metric of Year 1. In the context of the NEAT design, for MC items this is accomplished by giving Year 1 examinees and Year 2 examinees a common set of items. The Year 2 common item parameter estimates are adjusted to match as closely as possible the Year 1 estimates. This same adjustment is applied to the other parameter estimates from Year 2. In this way, adjustment is made for the difference in examinee ability across years so that the item parameter estimates will be on the same scale.

Implicit in Tate's (1999) argument is the notion that the common CR items are not really common, because different rating teams scored them. Thus, an extra step must be added to the usual equating process. In this step, adjustments to the Year 1 common item parameters are made that control for rater severity. This is accomplished by conducting a special linking study in which Team 2 raters score Year 1 responses on the CR common items. The Year 1/Team 2 CR and MC common items are then calibrated to obtain Year1/Team 2 common item parameter estimates. These estimates reflect the rating standards of Team 2 expressed in the metric of Year 1.

Once the Year 1/Team 2 common items have been placed on scale, the Year 2 items can be linked through the Year 2/Team 2 common items. The differences between the Year 1/Team 2 and the Year 2/Team 2 common item parameter estimates will reflect only the differences between Year 1 and Year 2 examinees, because the special linking study has effectively held constant the average rating team severity. Thus, by linking the two sets of parameter estimates, the Year 2 new form is correctly placed on the original scale of the Year 1 reference form as desired.

Tate (2003) and Kamata and Tate (2005) used simulation studies to show the effectiveness of the proposed IRT linking method incorporating trend scored papers. 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 methods) and nonlinear (e.g., frequency estimation, chained equipercentile) methods are often used to link CR tests. A review of the literature revealed no published study that examined the effectiveness of equating designs incorporating trend scoring (i.e., rescoring the same examinees to control for rater effects across scoring sessions, and comparing the two sets of ratings for the same group of examinees to eliminate group ability differences) in actual data from an operational test. Although there are operational programs that use equating designs incorporating trend scoring (e.g., some teacher licensure testing programs, NAEP), no study to date has examined non-IRT equating methods with trend-scoring designs.

The major purpose of the present study was to examine systematically some procedures currently used to link mixed-format tests in the context of the NEAT data-collection design. The study focused on classical equating methods. Two procedures did not use trend scoring: The first used only MC items in the anchor; the second used both MC and nontrend CR items. Two procedures incorporated trend scoring: One used essentially the procedure suggested by Tate (1999), adapted for non-IRT equating methods; the other used a hybrid of a single-group and equivalent-groups 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? (b) Which anchor test composition (MC and CR items, or MC items only) works best?

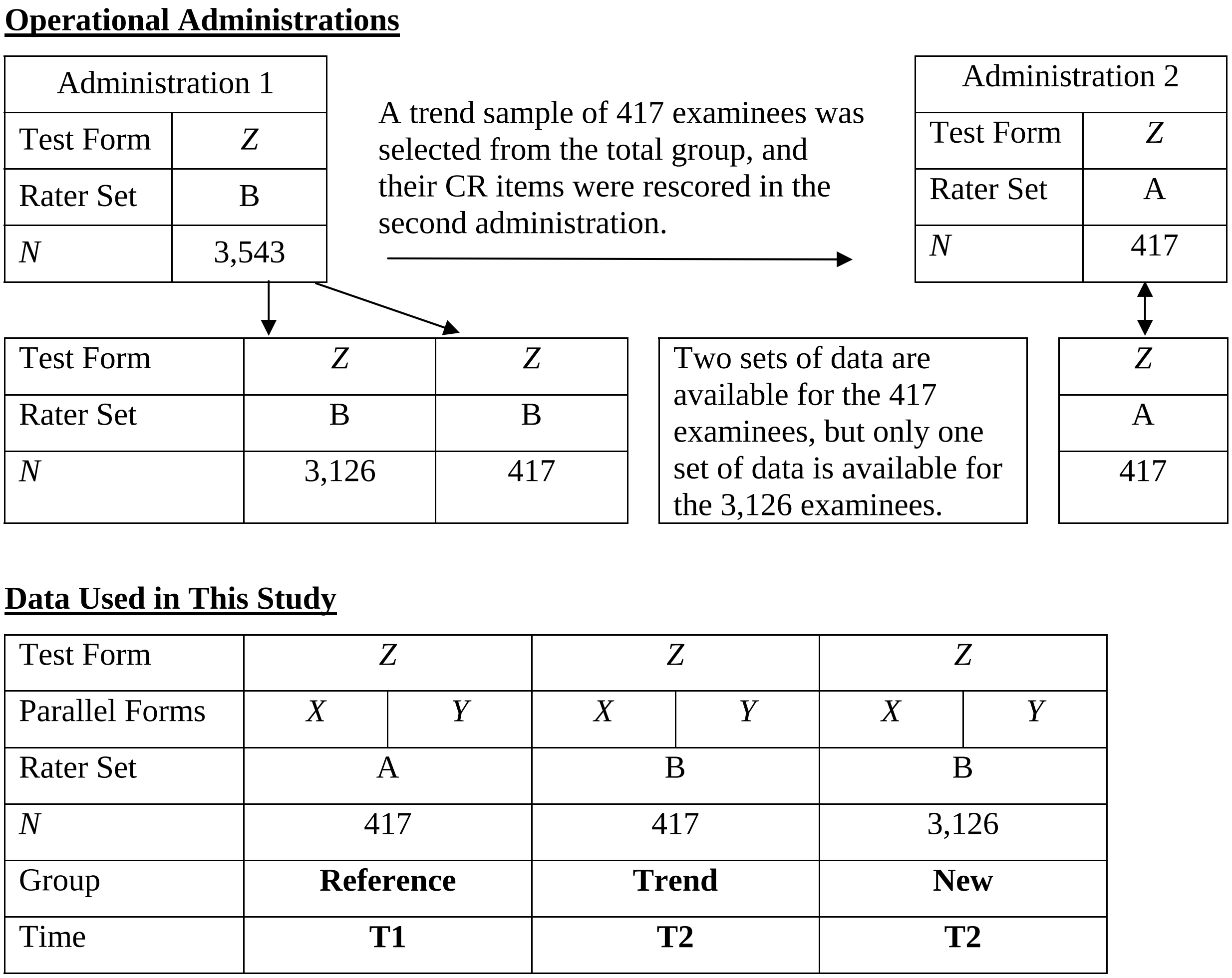
Method

Data

The data for the study were taken from a subject test of a large-scale testing program. This test consists of 24 MC items and 12 CR items. The possible score range of this test was 0 to 72, because each CR rating was an integer from 0 to 2 and all the CR ratings were weighted by 2. The data set from a single national administration of the test form was used. The sample size for the administration was 3,543 examinees (2,754 nonrepeaters, 789 repeaters). A sample of 417 examinees was randomly selected from the nonrepeater examinee group (N = 2,754), and the 12 CR items for this sample were re-scored by a different set of raters (a process known as *trend scoring*). As a result, two independent sets of scores for all CR items were available for those 417 examinees.

In an equating situation using trend scoring (e.g., see Tate, 1999), the CR items from the old form population are scored by a set of raters (call them Rater Set A). The CR items from the old form are re-scored by a different set of raters, the same raters that score the CR items for the new form (call them Rater Set B). To re-create this scenario in the present study, the 3,126 (= 3,543 – 417) examinees with single-scored CR items were treated as the new form population, and the 417 examinees scored by two sets of raters were treated as the reference form population. The set of raters who scored items for all 3,543 examinees were designated Rater Set B; whereas the set of raters who only scored items for the 417 examinees were designated Rater Set A. Figure 1 diagrams the data layout in this study. The figure also illustrates how the data, as used in this study, correspond to the actual administration from which the data were obtained.

One might question why the new form population in this study consisted of both nonrepeaters and repeaters although the reference form population consisted only of nonrepeaters.



Operational Administrations

Administration 1	
Test Form	*Z*
Rater Set	B
N	3,543

A trend sample of 417 examinees was selected from the total group, and their CR items were rescored in the second administration.

Administration 2	
Test Form	*Z*
Rater Set	A
N	417

Test Form	*Z*	*Z*
Rater Set	B	B
N	3,126	417

Two sets of data are available for the 417 examinees, but only one set of data is available for the 3,126 examinees.

Z
A
417

Data Used in This Study

Test Form	*Z*		*Z*		*Z*	
Parallel Forms	*X*	*Y*	*X*	*Y*	*X*	*Y*
Rater Set	A		B		B	
N	417		417		3,126	
Group	**Reference**		**Trend**		**New**	
Time	**T1**		**T2**		**T2**	

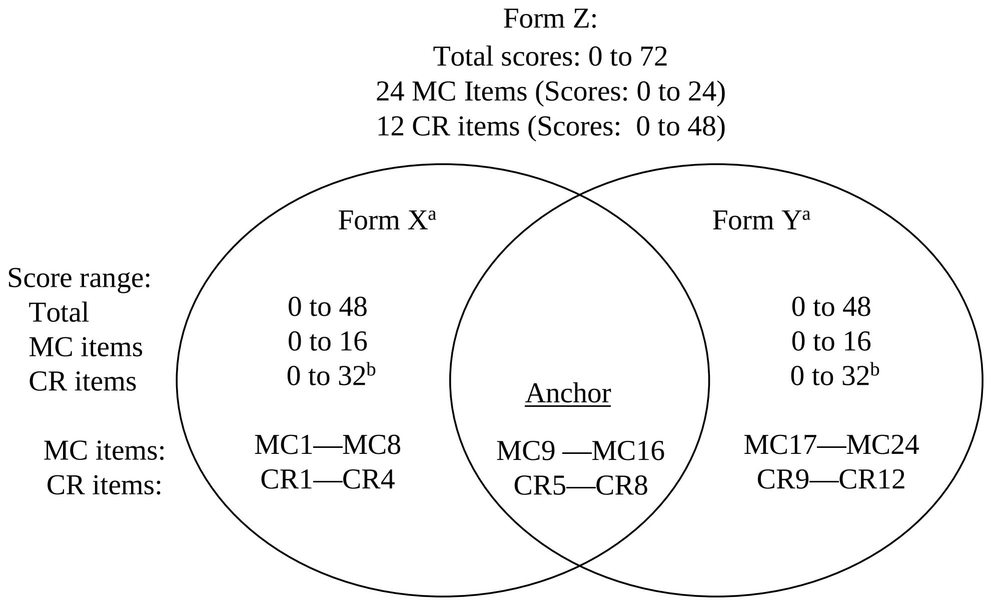
Figure 1. **General layout that presents how data were used in this study.**

Note. The 417 examinees were the reference group and the 3,126 examinees were the new group in this study. Two parallel forms, *X* and *Y*, were created from form *Z*, and they have 8 MC and 4 CR items in common.

The major purpose of this study was to assess different types of anchor designs in a situation where equating groups are not equivalent in ability (i.e., using a NEAT design). In this study, both reference and new form groups were drawn from a single administration, and thus they were relatively equivalent in ability, which is not always the case in reality. Quite often the groups being equated are not comparable for many reasons (hence, the NEAT design). Because repeaters tend to score lower than nonrepeaters, this study made use of all available examinees in the new form population regardless of their repeater status so that the new and old form groups would differ in ability as in operational testing situations.

Simulated Forms

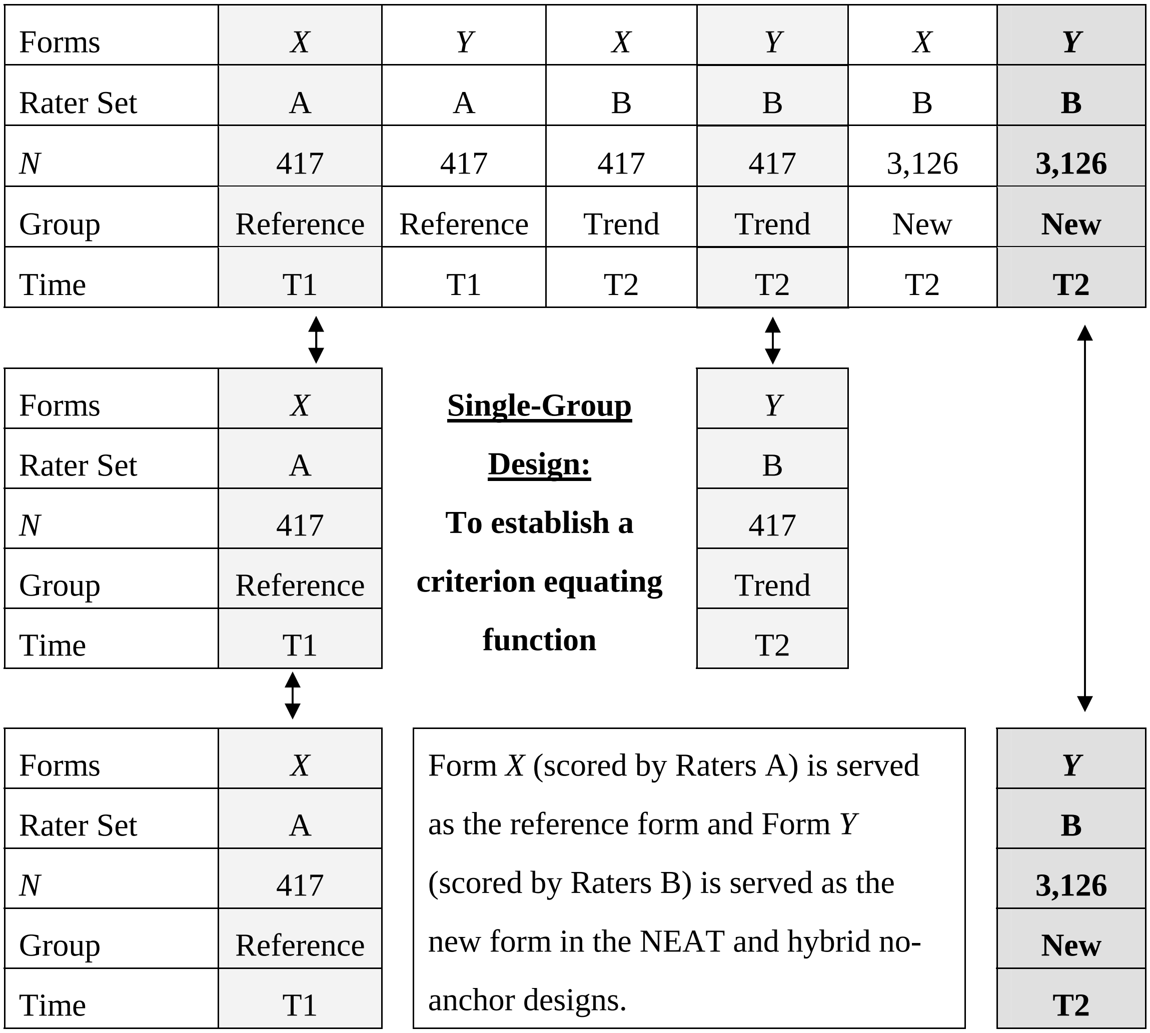
Two forms parallel in both content and difficulty (designated Forms *X* and *Y*), each comprising 16 MC and 8 CR items, were created from the original test (called Form *Z*), which contained 24 MC and 12 CR items. Figure 2 shows the basic layout for the two parallel forms. As shown, Forms *X* and *Y* have 8 MC and 4 CR items in common. The possible score ranges for the test and anchor were 0 to 48 and 0 to 24, respectively. The construction of two forms from a test given at a single administration allowed us to mimic the typical equating of alternate forms while having the advantage of obtaining data from a single group of examinees that took all of the items on both forms. For the purposes of the study, the scores of the 417 examinees whose CR items for Form *X* had been scored by Raters A served as the reference population. The scores of the 3,126 examinees whose CR items for Form *Y* had been scored by Raters B served as the new form population. Figure 3 presents the schematic of two parallel forms and equating groups.



a. The items 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.

Figure 2. **Design of the two simulated test forms used in the study.**



Forms	*X*	*Y*	*X*	*Y*	*X*	***Y***
Rater Set	A	A	B	B	B	**B**
N	417	417	417	417	3,126	**3,126**
Group	Reference	Reference	Trend	Trend	New	**New**
Time	T1	T1	T2	T2	T2	**T2**

Forms	*X*
Rater Set	A
N	417
Group	Reference
Time	T1

<u>Single-Group Design:</u>
To establish a criterion equating function

Y
B
417
Trend
T2

Forms	*X*
Rater Set	A
N	417
Group	Reference
Time	T1

Form *X* (scored by Raters A) is served as the reference form and Form *Y* (scored by Raters B) is served as the new form in the NEAT and hybrid no-anchor designs.

Y
B
3,126
New
T2

Figure 3. **Schematic of two parallel forms and equating groups, and the single-group design to establish the criterion for the study.**

Procedure

Criterion. The study examined ways to place the new form, Form *Y*, on scale with the reference form, Form *X*, using different designs and anchor compositions. The criterion represented the true linking of Form *Y* to Form *X* as shown in the middle section of Figure 3. For the 417 reference group examinees, two independent sets of scores for all CR items (Forms *X* and *Y*) scored by both Raters A and B were available. Accordingly, the true linking was estimated with those 417 examinees using a single-group design. To estimate the criterion function, total scores on Form *Y* (48 score points, Raters B) were equated to total scores on Form *X* (48 score points, Raters A) by using a mean-sigma equating method in the single-group design.

The data were also presmoothed using loglinear methods, and a direct equipercentile link was established to examine the curvilinearity of the relationship.

Equating designs. Two equating designs, (a) a NEAT design and (b) a hybrid no-anchor design, were considered in this study. In both designs, the reference form was Form *X* and the new form was Form *Y*. The 417 examinees were the reference form group and the 3,126 examinees were the new form group in both designs. The schematics of the NEAT and hybrid no-anchor designs are presented in Tables 1 and 2, respectively.

Table 1

Nonequivalent-Groups Anchor Test (NEAT) Designs With Three Different Anchor Sets

	Design 1A: MC plus no-trend CR anchor			
Group	Reference group[a]		New group[b]	
Score	Total	Anchor (internal)	Anchor (internal)	Total
Forms	*X*	*X*	*Y*	*Y*
N of MC	16	8	8	16
N of CR	8	4	4	8
Score range	(0–48)	(0–24)	(0–24)	(0–48)
CR raters	A	A	B	B
Time	T1	T1	T2	T2
	Design 1B: MC plus trend CR anchor			
Group	Reference group		New group	
Score	Total	Anchor (external)	Anchor (internal)	Total
Forms	*X*	*X*	*Y*	*Y*
N of MC	16	8	8	16
N of CR	8	4	4	8
Score range	(0–48)	(0–24)	(0–24)	(0–48)
CR raters	A	B	B	B
Time	T1	T2	T2	T2
	Design 1C: MC-only anchor			
Group	Reference group		New group	
Score	Total	Anchor (internal)	Anchor (internal)	Total
Forms	*X*	*X*	*Y*	*Y*
N of MC	16	8	8	16
N of CR	8	0	0	8
Score range	(0–48)	(0–8)	(0–8)	(0–48)
CR raters	A	NA	NA	B
Time	T1	T1	T2	T2

[a] *N* = 417. [b] *N* = 3,126.

Table 2

The Format of a Hybrid No-Anchor Design

	Design			
	Single-group		Equivalent-groups	
N	417	417	1,563	1,563
Score	Total	Total	Total	Total
Forms	*X*	*X*	*X*	*Y*
N of MC	16	16	16	16
N of CR	8	8	8	8
Score range	(0–48)	(0–48)	(0–48)	(0–48)
CR raters	A	B	B	B
Time	T1	T2	T2	T2

The first design, the NEAT design, is the one currently in use for this subject test. As mentioned previously, the reference population consisted of scores of the 417 examines whose CR items on Form *X* were scored by Raters A. The new form population was the 3,126 examinees whose CR items on Form *Y* were scored by Raters B. As presented in Table 1, Equating Design 1 was used with three different anchor compositions: (a) both MC and no-trend CR items, (b) both MC and trend CR items, and (c) only MC items. For the mixed-anchor cases [(a) and (b)], Form *Y* was linked to Form *X* through the common items, for which common CR items had been scored by the same raters (the trend CR case) or different raters (the no-trend CR case) across the reference and new form groups.

In the Design 1A case, four common CR items were scored by different sets of raters across the reference (by Raters A) and new (by Raters B) form groups. Because the trend-scoring information was not utilized to adjust for any scoring shift over time, the success of equating rested on the assumption that Raters A and B used the same scoring standards. In this case, the common CR scores represented internal anchors, along with common MC scores, because the anchor scores formed part of the total scores. As shown in Table 1, Design 1B had exactly the same format as Design 1A. In Design 1B, however, the four common CR items were scored by the same Raters B in both the reference and new form groups. Because the CR anchor items for the 417 reference examinees were rescored by Raters B together with the 3,126 new examinees using a trend-scoring method, any CR scoring shift caused by different sets of raters in the reference (Raters A) and new (Raters B) form groups could be adjusted. In the reference form group, the CR scores of the anchor were external, because they were generated at a different time point by different raters compared to the total scores. In the new form population, however, the CR scores of the anchor were internal because they were generated at the same time point by the same raters with the CR scores of the total test.

The second design, called the *hybrid no-anchor design*, represented an alternative to the NEAT design. As displayed in Table 2, this design was a combination of a single-group design (reference form group) and an equivalent-groups design (new form group). The hybrid no-anchor design would be possible if a reference form was 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 equivalent-groups design. As the name indicates, an anchor test (i.e., common MC items, CR items, or both over two spiral forms) is not necessary in this design, although the use of anchor items across the two spiral forms could enhance the accuracy of equating functions. The specific linking procedures in the hybrid design were as follows.

For the hybrid design, in the reference group, Form *X* scored by Raters 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) was linked to Form *X* scored by Raters A using a single-group design (*N* = 417). In the new form group, the 3,126 examinees were randomly split between the new Form *Y* (*N* = 1,563) and the reference Form *X* (*N* = 1,563). Then Form *Y* scored by Raters B (*N* = 1,563) was linked to Form *X* scored by the same Raters B (*N* = 1,563) using an equivalent-groups design.

Evaluation

For all equating designs, linear equating methods (e.g., chained linear, Tucker, and Levine methods) were used. Many observed-score equating methods are based on the 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 Form *Y* equated raw scores obtained using each equating method in each equating design were compared with the criterion. The differences among the conversions were quantified using the root mean squared difference (RMSD),

$$RMSD = \sqrt{\sum_{i=0}^{48} w_i \left[\hat{e}_i\left(x_i\right) - e_i\left(x_i\right) \right]^2} , \qquad (1)$$

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 an 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 hybrid design) form groups. Then Form *Y* scores were equated to Form *X* scores 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 equating error variance at each score point, and the square root of this value defined the conditional RMSE index. The following equations represent bias, equating error (CSEE), and RMSE measures conditioned on each raw score point (x_i):

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

$$SEE_i = s(d_i) = \sqrt{Var_j\left[\hat{e}_j\left(x_i\right) - e\left(x_i\right)\right]} = \sqrt{Var_j\left[\hat{e}_j\left(x_i\right)\right]}. \qquad (3)$$

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

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

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, $\sqrt{\sum_i w_i SEE_i^2}$; 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 Form *Y* were equated to total test scores on Form *X* with a total of 417 examinees based on a single-group design to define the criterion. As shown in Figure 3, CR scores on Forms *X* and *Y* were generated by Raters A and B, respectively. The means and standard deviations were 33.31 and 5.96 for Form *X* and 33.68 and 5.97 for Form *Y*, respectively. For each raw score on Form *Y*, the equivalent raw scores on Form *X* were determined using the mean-sigma (linear) and direct equipercentile (nonlinear) methods. Figure 4 presents equated raw score differences between the mean-sigma and direct equipercentile methods. For almost all raw score points (0 to 48), the differences between the two functions were less than the *difference that matters* (Dorans & Feigenbaum, 1994), defined as half of a raw score point. Because the differences between two equating functions were considered negligible, 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 seems to be reasonable in that the criterion functions were derived from a relatively small sample (*N* = 417) and linear equating methods were used in all the equating designs.

Anchor Design

Summary statistics of total and anchor scores for examinee groups taking Forms *X* and *Y* are presented in Table 3. The total score mean of the Form *X* group (*M* = 33.31) was higher than that of the Form *Y* group (*M* = 31.32). Regarding the anchor, the Form *X* group showed higher means for both anchor formats than the Form *Y* group showed, because the Form *X* group consisted of only first-time test takers who tended to score higher than test repeaters. The effect size of the difference between the anchor means was .31 in Design 1B and .16 in Design 1C. This magnitude indicates a fairly large difference in ability between the two groups for this type of testing program. In both cases, the Form *X* group was more proficient than the Form *Y* group.

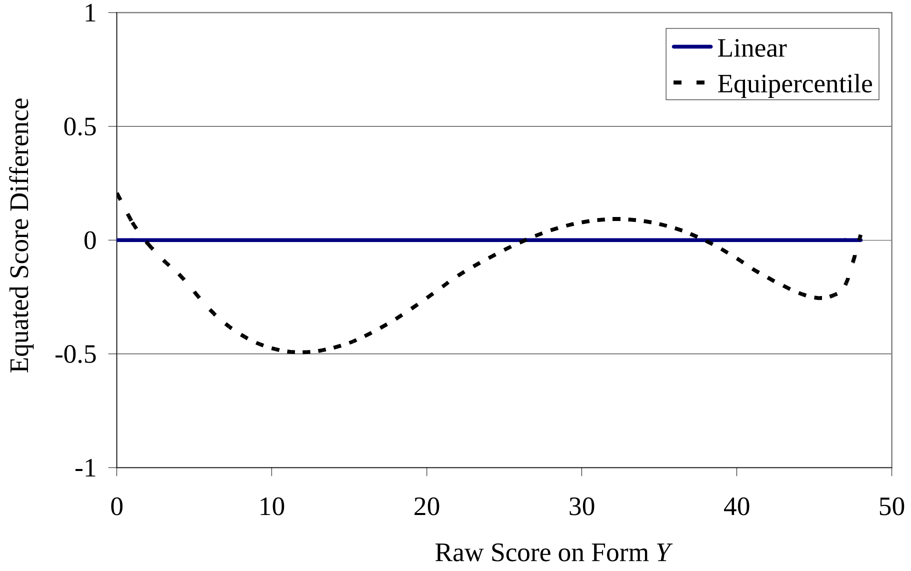


Figure 4. **Difference plot between linear and direct equipercentile criterion functions.**

Table 3

Summary Statistics for Examinee Groups Taking Forms X and Y in the Anchor Design

	Reference form *X*		New form *Y*	
	Total	Anchor	Anchor	Total
Sample size	417		3,126	
Design 1A: MC plus no-trend CR anchor				
Number of items	24	12	12	24
Mean	33.31	16.17	15.66	31.32
Standard deviation	5.96	3.57	3.85	6.92
Anchor-total corr.	0.86		0.87	
Design 1B: MC plus trend CR anchor				
Number of items	24	12	12	24
Mean	33.31	16.80	15.66	31.32
Standard deviation	5.96	3.41	3.85	6.92
Anchor-total corr.	0.69		0.87	
Design 1C: MC-only anchor				
Number of items	24	8	8	24
Mean	33.31	5.98	5.77	31.32
Standard deviation	5.96	1.26	1.34	6.92
Anchor-total corr.	0.55		0.57	

In the MC-only anchor test design, the magnitude of the correlations between the total test score and the MC-only anchor scores were relatively low (*r* = .55 –.57) but fairly similar in both groups. The magnitude of the correlations between the total test score and MC plus no-trend CR anchor scores were high (*r* = .86 – .87) and very similar in both groups. However, the mixed anchor based on trend CR items was correlated more highly with the total score in the Form *Y* group (*r* = .87) than in the Form *X* group (*r* = .69); here the CR anchor was internal for the Form *Y* group but external for the Form *X* group. As explained previously, using a trend-scoring method, the CR anchor scores were generated at a different time by a different set of raters (Raters B) than the CR score part of the total scores in the reference form group. Accordingly, the CR anchor scores, called *external* here, were not part of the total test scores.

Table 4 presents the difference between each linear equating function and the criterion, using the RMSD deviance measure. Figure 5 plots the conditional equated score difference between the chained linear equating function and the criterion in each equating design. 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 adjusts as if the equating samples were more similar in ability than the anchor scores would indicate. As a result, we would expect Tucker equating to be biased, because Form *X* and Form *Y* groups differed substantially in this study.

Table 4

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	Chain linear	Tucker	Levine
NEAT: MC plus no-trend CR anchor	1.593	1.743	1.551
NEAT: MC plus trend CR anchor	0.414	1.048	0.178
NEAT: MC-only anchor	1.490	2.003	0.926
Hybrid no-anchor	0.129	--	--

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 this mixed-anchor case than in both the MC-only anchor and MC plus 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. This result clearly indicated the potential problems caused by using no-trend CR anchor in a NEAT design.

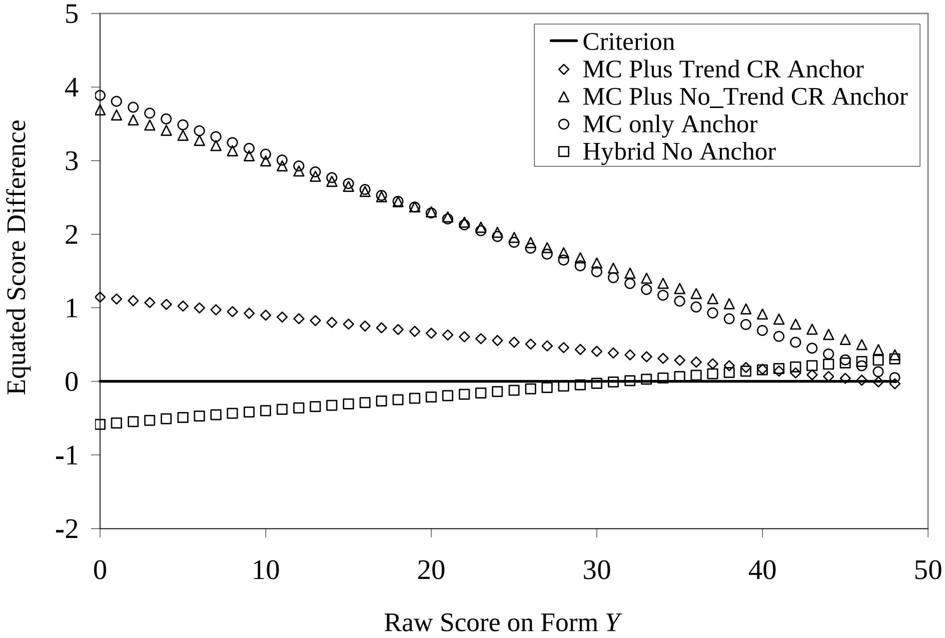


Figure 5. **Equating results of the chained linear method in the four equating designs.**

Table 5 presents the summary of the weighted average root mean squared bias, equating error, and RMSE derived from a boostrap resampling technique for each equating design with chained linear equating. Again we chose the mean-sigma results from the single-group design as the equating criterion when examining equating bias and error based on the resampling technique. As shown, using a no-trend CR anchor along with MC resulted in the largest bias but the smallest error. The use of MC items alone resulted in much larger bias (as expected), but also slightly larger equating error than resulted from the MC plus trend CR anchor case. As expected, the MC plus trend CR anchor yielded the smallest bias compared to the other two anchor design cases. Although equating error was fairly comparable for those three designs, the magnitude of bias was substantially larger in both the MC-only and the MC plus no-trend CR anchor cases than in the MC plus trend CR case.

Figures 6 to 8 plot the conditional bias of chained linear equating in the anchor design, along with an error band representing plus or minus one empirical conditional standard errors of equating. The error band for chained linear equating was slightly wider in the MC-only anchor case than in the other mixed-anchor cases, implying severe fluctuation across 500 replications, particularly for the raw score range of 0 to 20. The results indicate that the MC plus trend CR anchor would be better than the MC-only anchor for the mixed-format tests in this case.

Table 5

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: MC plus no-trend CR anchor	1.603	0.238	1.620
NEAT: MC plus trend CR anchor	0.415	0.360	0.549
NEAT: MC-only anchor	1.496	0.420	1.554
Hybrid no-anchor	0.084	0.401	0.410

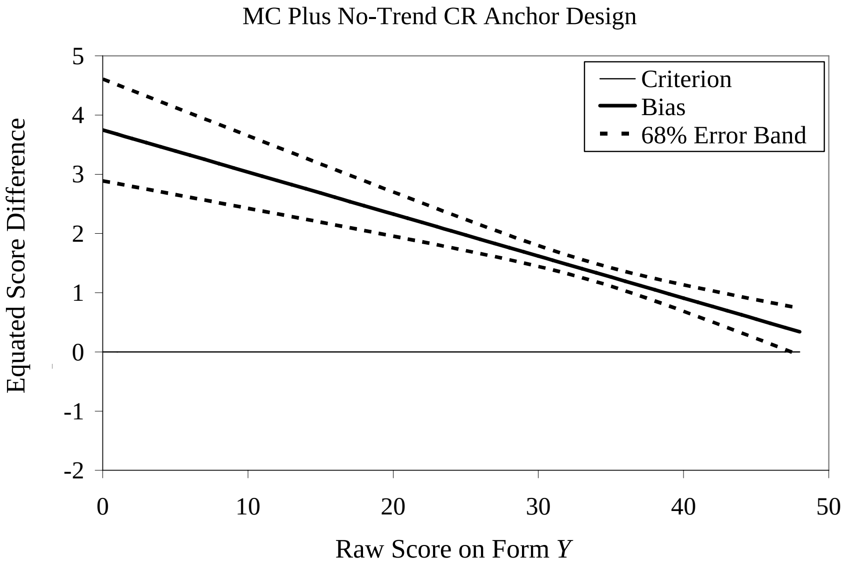


Figure 6. **Difference of equating results from the criterion, for the nonequivalent-groups anchor test design with MC plus no-trend CR items in the anchor.**

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

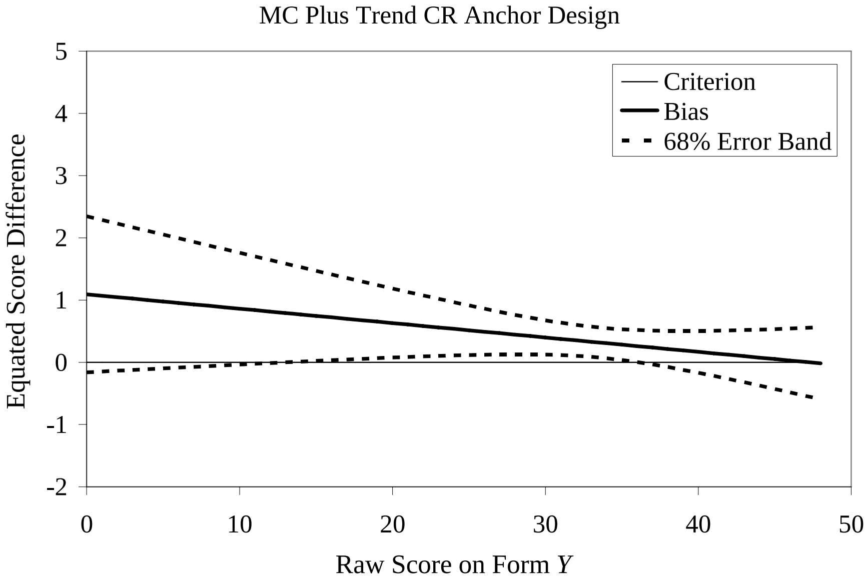


Figure 7. **Difference of equating results from the criterion, for the nonequivalent-groups anchor test design with MC plus trend CR items in the anchor.**

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

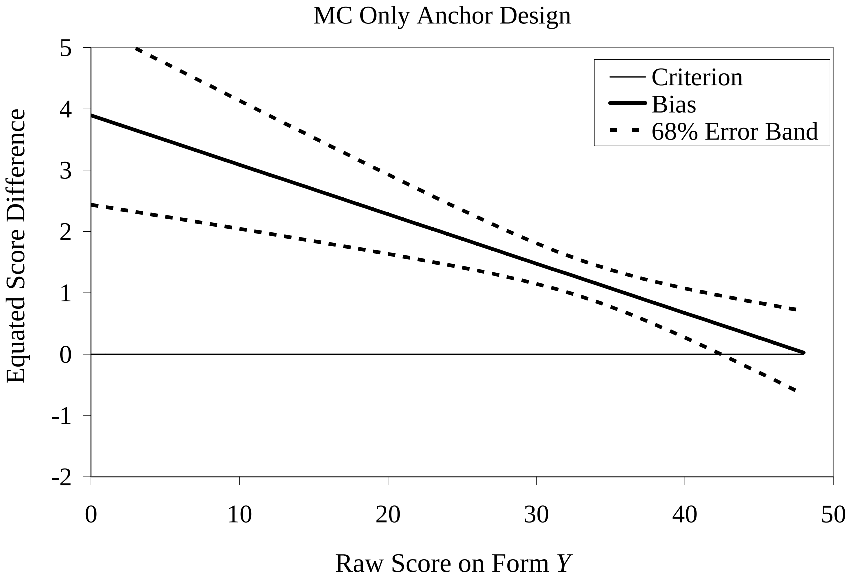


Figure 8. **Difference of equating results from the criterion, for the nonequivalent-groups anchor test design with only MC items in the anchor.**

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

Hybrid No-Anchor Design

Summary statistics of the total scores for examinee groups taking Forms *X* and *Y* are presented in Table 6. In this design, the new form sample (*N* = 3,126) was randomly split to simulate the spiraling of the new Form *Y* with the reference Form *X*. Because Forms *X* and *Y* are nearly parallel, the total means are almost identical for both groups. As shown in Figures 1 and 3, for the 417 examinees who took Form *X*, two sets of total scores are available, because two different sets of raters, Raters A and B, scored their CR items at different times. Form *X* scores for the reference sample (*N* = 417) and for the new form sample (*N* = 1,563) can be directly compared, because the same Raters B generated scores for the CR items in both cases. As expected, the 417-examinee group was more proficient than the 1,563 examinees taking either Form *X* or Form *Y*.

Table 6

Summary Statistics for Examinee Groups Taking Forms X and Y in the Hybrid No-Anchor Design

Group	Reference	Reference	New	New
Form	*X*	*X*	*X*	*Y*
CR Raters	A	B	B	B
Sample size	417	417	1,563	1,563
Number of items	24	24	24	24
Mean	33.31	33.95	31.48	31.35
Standard deviation	5.96	6.06	7.10	6.91

The summary statistics for the hybrid no-anchor design are also summarized in Tables 4 and 5, along with the results for the anchor designs. Although there was no anchor test for the hybrid design, the equations used to obtain the linear equating results are indistinguishable from the equations for the chained linear equating in the NEAT design case. As shown in Table 2, the first half of the chain linked the total scores on Form *X* generated by Raters B to the total scores on Form *X* generated by Raters A for the 417 examinees using a single-group design. The second half of the chain linked the total scores on Form *Y* generated by Raters B to the total score on Form *X* generated by Raters B using an equivalent-groups design. Although in the first link each examinee had both scores (one derived from the scoring by Raters A and the other from Raters B), in the second link each examinee had only one score—from either Form *X* or Form *Y*.

Consequently, 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 this case. For that reason, only the results from the chained linear method were examined for the hybrid design.

The hybrid no-anchor design yielded the smallest RMSD of the four designs. In general, the hybrid no-anchor design appeared to fare well with respect to bias and equating error. The hybrid design resulted in the smallest bias, which is promising especially for tests with cut scores. Although the equating error was slightly larger than for the anchor test design with MC and trend CR anchor items, the size of equating error was quite small in all designs. Overall, the hybrid design resulted in the smallest RMSE value, and the anchor design using MC plus trend CR anchor items yielded the next smallest RMSE value. Interestingly, the mixed anchor with no trend CR scoring yielded the largest RMSE value. Figure 9 plots the conditional bias of chained linear equating in the hybrid design, along with an error band representing plus or minus one empirical CSEE. The conditional bias was negligible across all the raw score points.

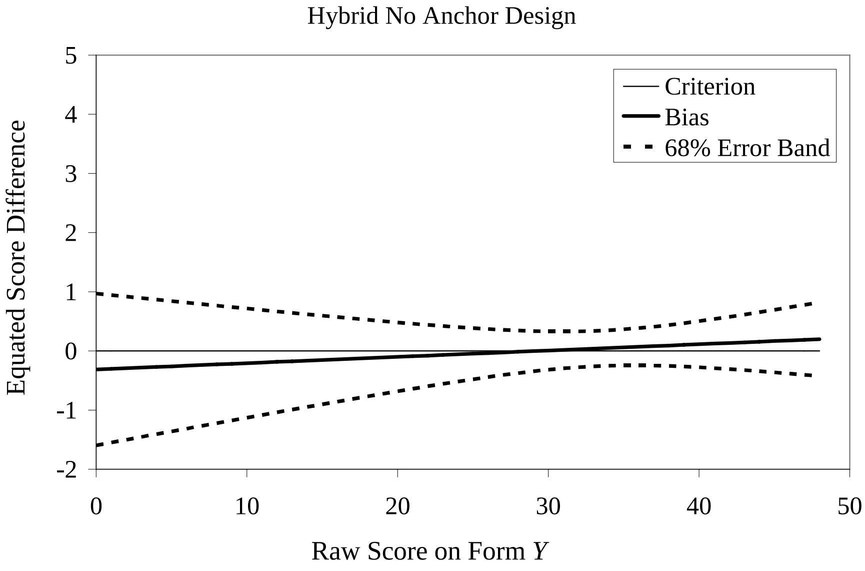


***Figure 9.* Difference of equating results from the criterion, for the hybrid single-group/equivalent-groups design.**

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

Conclusions

The present study examined systematically some possible methods for linking CR tests in an attempt to discover which designs are most effective in adjusting CR items and tests for difficulty mostly caused by rater severity. Different equating designs were compared for a test with a mix of MC and CR items. Many methods of test equating have been proposed and well documented (Kolen & Brennan, 2004). These methods, however, often are not useful for tests with CR items, because the methods do not make allowances for scoring inconsistency peculiar to CR items. For tests that use CR items, scoring consistency over time should be investigated to ensure the accuracy of examinees' scores. If scoring standards do in fact shift, application of the standard linking or equating methods or designs developed for MC item tests may yield inaccurate results, because changes in scores on CR anchors due to a scoring shift will be inappropriately attributed to ability differences of the examinee groups (Tate, 1999).

Many practitioners may overlook the differences in application of CR scoring standards across test form administrations and attempt to use conventional equating methods. The use of the mixed anchor might be harmful, however, when no-trend CR items are incorporated as an anchor in the presence of a change in CR scoring standards. In this study the MC plus no-trend CR anchor design showed the greatest amount of bias among the four designs employed in this study, providing clear evidence of the danger of not taking into account possible changes in scoring standards. Common CR items should therefore be used as anchor items with caution. 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 a cut score, the use of traditional equating methods using no-trend scored CR anchors could result in incorrect pass/fail decisions for examinees.

As mentioned previously, some practitioners have suggested using MC items as anchors to control for differences among test forms containing CR items (e.g., see Baghi et al., 1995; Ercikan et al., 1998). This format of equating may be inappropriate, though, due to the possible multidimensionality of mixed-format tests (Bennett et al., 1991; Sykes et al., 2002). Previous research showed that use of MC-only anchors could result in potentially large equating bias (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999). This research is consistent with those previous findings. The MC-only anchor design produced large RMSD, bias, equating error, and RMSE, showing inferiority to the mixed-anchor (MC plus trend CR) and hybrid no-anchor designs.

This study showed that equating bias caused by a scoring shift could be controlled using a trend-scoring method in practice. The trend-scoring method could be expensive and possibly difficult to implement; however, in an image or online scoring system with proper tools, it would be straightforward to implement in practice. The trend-scoring method has statistical strengths in detecting a CR scoring shift. As shown in this study, the trend CR anchor displayed much better performance than did the no-trend CR anchor in recovering the true equating function, primarily because of the bias reduction. The equating error was actually slightly higher than that for the MC plus no-trend CR anchor. This may be attributable in great part to the somewhat lower anchor (external) total correlation for the reference form 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 mixed anchor when trend scoring was used.

In this study, the MC plus trend CR anchor design displayed superior performance to the other two anchor designs, but this design yielded larger bias and RMSE values than the hybrid (possibly no anchor) design produced. Among the four designs, the hybrid design seems the best model psychometrically in adjusting for changes in the scoring standards for the CR common items. The present research shows that spiraling two forms (e.g., new and reference) may offer psychometric advantages over the current practice (i.e., a NEAT design). In general, the superiority of the random groups design over the NEAT design is that the representativeness of the anchor becomes irrelevant because the anchor is unnecessary in the random-groups design. Even so, the use of a properly constructed anchor across the two spiral forms in the random-groups design (an equivalent-groups anchor test design) will certainly improve the accuracy of equating functions.

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

NEAT design to achieve the same level of equating error. Table 7 summarizes the practical comparisons between the two designs.

Table 7

Practical Comparisons Between MC Plus Trend CR Anchor Design and a Hybrid No-Anchor Design

Aspects	MC plus trend CR anchor design	Hybrid no-anchor design
Common items in forms	Required	Not required
Rescoring CR items	Common items only	All CR items
Administration	New form only	New and reference forms
Large sample size	Maybe	Required

Given the limitations listed above, practitioners may choose one or the other of the NEAT or hybrid designs, depending on the situation. The NEAT design may be preferred when the number of CR items to be scored must be kept as low as possible; when sample sizes in each administration are relatively small, such that spiraling would result in insufficient numbers of examinees to insure random equivalence or sufficiently small equating error; or when security issues or other concerns preclude re-administering the reference form. The hybrid 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 hybrid design might also allow equating of short all-CR tests or other tests in which no CR items are reused.

The present study is meaningful for two reasons. First, this study examined the effectiveness of equating designs incorporating trend scoring using non-IRT equating methods, whereas previously published studies have examined IRT-based equating methods only. Second, previous studies have involved primarily simulated data, whereas this study applied the methodology to actual data from an operational test. The findings of the present study are promising, and thus many practitioners may consider them when designing their CR test equating.

The results of this study also draw attention to a third issue, often overlooked in operational settings. In many cases, a test form may be reused, and the original test score conversion obtained when the test was first equated is applied in subsequent administrations.

This research demonstrates, however, that if the test contains CR items, the original test score conversion may no longer apply. The reason is that the scoring standards may not be constant across all 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 test form should be re-equated to adjust for differences in rater severity.

There may be limitations in generalizing the findings of the current study. This study is based solely on a single test form with a single administration. The criterion was derived from relatively small samples (less than 500). Additional empirical evidence about the hybrid design should be gathered using various data sets from different formats, different subject tests, and different administrations to enhance its generalizability. In this application, Forms *X* and *Y* have 8 MC and 4 CR items in common, by design. It is generally assumed that equating will be effective when two test forms are close to parallel, which was the case in this study. The effectiveness of the hybrid no-anchor design should be evaluated in a situation where the spiraled forms are not parallel.

For a test with passing scores, equating accuracy is much more important in the passing score range than at other score points. For assessments in which a pass/fail decision is rendered based on the total test score, the focus should be on the standard errors of equating at and near the cutoff score. Accordingly, one way to assess the benefits of each equating design is to examine its impact on examinees with respect to pass/fail decisions. Deviance measures need to be calculated only for the cut-score region.

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.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*, 1210–1218.

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

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

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

Donlon, T. F. (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Dorans, N. J., & Feigenbaum, M .D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10, pp. 91–122). Princeton, NJ: ETS.

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement*, *35*, 137–154.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, *14*(1), 31–57.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington, DC: National Center for Educational Statistics.

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement*, *42,* 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*, 357–381.

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, Canada.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 12–23.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Educational Statistics.

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*, 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,* 893–914.

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118.