Linking Mixed-Format Tests Using Multiple Choice Anchors

Michael E. Walker
Sooyeon Kim
ETS, Princeton, NJ

Paper presented at the annual meeting of the
American Educational Research Association (AERA) and the
National Council on Measurement in Education (NCME)

April 13-17, 2009, San Diego, CA.
Abstract

This study examined the use of an all-multiple-choice (MC) anchor for linking mixed format tests, containing both MC and constructed-response (CR) items, in a non-equivalent groups design. A MC-only anchor could effectively link two such test forms if either: (a) the MC and CR portions of the test measured the same construct, so that the MC anchor adequately represented the entire test; or (b) the relationship between the MC portion and the total test remained constant across the new and reference linking groups. The study also evaluated whether linking mixed-format tests through MC-only anchors would be more effective than would a two-stage strategy in which MC portions were equated through MC anchors and then composite scores were scaled to the MC scores. Anchor linking and two-stage linking yielded identical (or nearly so) results for both linear and nonlinear linking methods. The paper discusses some advantages of a direct approach.
**Introduction**

Many large-scale testing programs include constructed-response (CR) as well as multiple-choice (MC) items in their assessments. As with other standardized tests, these mixed-format tests must be equated to insure equivalence of scores across test forms. Perhaps most often, equating occurs in the context of the non-equivalent groups with anchor test (NEAT) design, in which a set of items common to both the new and reference forms is used to place both forms on the same scale. These common items should represent the entire test form in terms of content and difficulty.

NEAT equating has proven difficult with mixed-format tests. One reason is that, because CR items tend to be easy to memorize (Muraki, Hombo, & Lee, 2000), it may be difficult to find CR items that can be reused across forms. Second, even if the same CR items are used, the standards of the raters scoring the CR items almost always differ across the two administrations (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). In this case, the CR anchor items would confound differences in rater severity with true group differences, so that adjustment through the anchor would lead to erroneous results (Tate, 1999).

One solution to the second problem involves rescoring CR papers from the reference form administration (a process called trend scoring). For a random group of examinees from the previous administration, the anchor CR items are rescored by the raters who score the new form in the current administration. The scores from the new raters (not the old) on the reference form become part of the anchor score. This method effectively nullifies any differences in scoring standards across administrations, because the same raters score the anchor items on both sets of papers. Equating with trend scoring
has proven successful in producing accurate linkings (Kim, Walker, & McHale, 2008; Tate, 1999, 2000). However, the time, logistic, and monetary costs of this method could be prohibitive in certain situations.

Another possible solution to the problem of lack of an appropriate CR anchor would be to attempt to link mixed format tests through MC items only. Indeed, some practitioners have suggested using MC items as anchors to control for differences 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 can lead to biased equating results (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999); possibly because MC and CR items measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002).

This research focuses on the use of an internal (i.e., part of the total test) MC-only anchor to place two mixed-format test forms on the same scale. There are at least two different ways we can envision linking through MC-only anchors. The first possibility is that if the MC part of the test measures the same construct as the CR part, and if the CR portions of the test are fairly consistent from form to form with respect to content and difficulty, then adjusting for group ability through a MC-only anchor would result in what could be called equating in the sense that the composite score for the new test would be aligned along the same dimension as the composite score for the reference test.

A second possibility is that the MC part of the test measures a somewhat different construct from the total test; or that the CR portions vary from form to form with respect to content and rater standards. In the context of the NEAT design, we could use scale alignment techniques to place the reference and new test scores on a common scale.
(Holland & Dorans, 2006). To do this, we could equate the MC portions of the reference and new test forms using the MC anchor. We could then scale both composite (MC plus CR) scores to the MC scores, so that both composite scores would be on the common MC score metric. Note that once both MC portions have been equated, the MC scores on the two tests may be considered equivalent. Thus, linking the composite scores in this way is an example of what Holland and Dorans (2006) call scaling to an anchor. Even if the MC and CR items measured somewhat different constructs, this method would still result in comparable scores, as long as the relationship between the MC and CR sections was identical across the reference and new examinee populations, leading to constant dimensionality across populations. This research examined whether either of these conditions (unidimensionality vs. constant dimensionality) held, such that mixed-format tests could successfully be linked through MC-only anchors.

Another purpose of the study was to evaluate a claim (Michael Kolen, personal communication, November 26, 2007) that linking mixed-format tests through MC-only anchors would result in something closer to equating than would a strategy in which MC portions were equated through MC anchors and then composite scores were scaled to the MC scores. One could counterclaim that the MC portions of the tests are properly equated using the MC anchor, whereas the composite scores are not. Furthermore, the relationship between the MC portion and the composite should be stronger than that between the MC anchor and the composite by virtue of the longer length of the entire MC section. Thus, we might expect scaling to an anchor to align the two test scores better than linking through the MC-only anchor.
A linking criterion

To better understand the relationship between anchor equating and scaling to an anchor in this context, it is worthwhile to investigate what the linking criteria should be for the two strategies. In the more direct anchor equating, the new composite is placed on the scale of the reference composite via the MC anchor tests. Figure 1A illustrates this procedure. In the figure, all arrows lead from the new composite to the reference composite. The goal of linking here is to estimate what the new-to-reference score relationship is in the population. Theoretically, the criterion would be found by directly linking the new scores to the reference scores in a population in which every individual had taken both tests.

With scaling to an anchor, illustrated in Figure 1B, the new MC portion is placed on the scale of the reference MC portion. Then both composites are placed on this reference MC scale through direct scaling. In the figure, all arrows lead to the reference MC portion. The criterion for such a linking would be on the scale of the reference MC portion.

The methods used in equating and scaling are symmetric by nature (see Kolen & Brennan, 2004; Holland & Dorans, 2006). This means that the function mapping one set of scores onto the other is one-to-one and therefore invertible. If we take the arrow leading from the reference composite to the reference MC portion in Figure 1B and reverse it, we get Figure 1C. The relationships implied by Figures 1B and 1C are equivalent, but they are on different scales. The scale in Figure 1C is that of the reference composite score, the same as in Figure 1A. As in Figure 1A, all arrows in Figure 1C lead from the new composite to the reference composite. The figure would imply that the
criterion for Figure 1C is the same as for Figure 1A, and is found by directly linking the new composite scores to the reference composite scores in a population in which every individual has both scores. Thus, Figure 1C represents a two-stage linking strategy whose goal is equivalent to the more direct anchor linking method depicted in 1A.

A. Anchor Linking

![Diagram for Anchor Linking]

B. Scaling to an Anchor

![Diagram for Scaling to an Anchor]

C. Two-Stage Linking

![Diagram for Two-Stage Linking]

FIGURE 1. Three strategies for linking the new composite to the reference composite using MC-only anchors.

We can illustrate this point algebraically as well using the equipercentile linking function. To do so, we adopt the notation of Braun & Holland (1982). Table 1 shows the symbols we will use for the scores involved in forming the criteria: the variable name and the associated cumulative distribution function (cdf) for each. Note that the MC-only anchor tests are not included in the table, because these would theoretically not be needed when forming the criterion in the population.
Table 1

*Symbols Used to Represent Composite and Multiple Choice Scores for Reference and New Forms in Illustrating Equivalence of Linking Criteria for Direct and Two-Stage Linking*

<table>
<thead>
<tr>
<th>Score</th>
<th>Symbol</th>
<th>cdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref Composite</td>
<td>$Y$</td>
<td>$G(y)$</td>
</tr>
<tr>
<td>Ref MC</td>
<td>$Z$</td>
<td>$H(z)$</td>
</tr>
<tr>
<td>New MC</td>
<td>$W$</td>
<td>$E(w)$</td>
</tr>
<tr>
<td>New Composite</td>
<td>$X$</td>
<td>$F(x)$</td>
</tr>
</tbody>
</table>

Assume that the scores in Table 1 are continuous random variables. Then, using the symbols from the table, the direct equipercentile function linking new composite ($X$) scores to reference composite ($Y$) scores may be written (Braun & Holland, 1982),

$$e_y(x) = G^{-1} \left( F(x) \right). \quad (1)$$

In words, the function that places new score $X$ on the scale of old score $Y$ is a compound function involving the inverse cdf of $Y$ and the cdf of $X$. Likewise, the two-stage linking strategy may be expressed as,

$$e_y(x) = G^{-1} \left( H^{-1} \left( E \left[ E^{-1} \left( F(x) \right) \right] \right) \right). \quad (2)$$

Cancellation of functions with their inverses leaves a result identical to Equation 1. In other words, both the direct and two-stage strategies yield the same criterion in this case.

The current research examined both direct anchor linking and a two-stage process to assess the effectiveness of an MC-only anchor in linking mixed format tests. Both
linear and nonlinear chained linking methods were used. We hypothesized that the use of a MC-only anchor would be effective to the extent that MC-composite relationship remained constant across the reference and new form groups; and that the direct and two-stage strategies would yield identical results in the chained linking case.

Method

Data

The data for the study were taken from two administrations of a subject test, comprising 24 MC and 12 CR items, of a large scale testing program. Each MC item received a maximum score of 1, and each CR item received a maximum score of 4. Thus, the possible score range of this test (called Form Z) was 0 to 72.

For one administration, the 12 CR items for 417 examinees were scored by Rater Group A. These 417 examinees constituted the reference group in this study. In another administration, the same 12 CR items for those 417 examinees were independently scored by Rater Group B. These same Raters (Group B) also scored the 12 CR items for a separate group of examinees ($N = 3,126$). These 3,126 examinees constituted the new group in this study. Note that two independent sets of scores for all CR items were available for the 417 reference examinees, but only a single set of CR scores was available for the 3,126 new examinees.

Simulated Forms

The original test form (Form Z) used in this study had 24 MC and 12 CR items. Two forms parallel in both content and difficulty (designated new form and reference
form) were created from Form Z. The new and reference forms consisted of 16 MC and 8 CR items. Those forms had 8 MC and 4 CR items in common. The common MC items were used as the anchor in a NEAT design. The maximum possible scores for the test and anchor were 48 and 8, respectively. For the purposes of the study, the reference form as scored by Rater Group A (reference form/Rater A) served as the reference form and the new form as scored by Rater Group B (new form/Rater B) served as the new form\(^1\).

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 yielding data from a single group of examinees that took all of the items on both forms. Because all examinees took both new and reference forms, the two test forms could be directly equated using a single group design, and the result could be used as a criterion to examine the effectiveness of the two linking strategies.

**Procedure**

*Criterion.* The criterion represented the true linking of the new form to the reference form. This linking was estimated using a single group design with those 417 examinees who took both the new and reference forms. As illustrated previously, this single criterion was used to evaluate both the direct anchor linking and the two-stage linking. The schematic of this design is presented in the upper section of Figure 2.

To estimate the criterion function, total scores on the new form were equated to total scores on the reference form by setting means and standard deviations equal in a

---

\(^1\) This is the typical situation for a test with CR items. Each administration involves a different set of raters. The advantage of the data in the current study was the availability of scores for the reference group from both sets of raters.
single group design. This criterion was used for the linear equating methods. The data were also pre-smoothed using loglinear methods, and a direct equipercentile link was established to produce a nonlinear criterion to use for the nonlinear linking methods.

Equating Strategies. Two linking strategies were used in the context of the NEAT design: (1) traditional chained linking of new to reference form using a MC-only anchor; and (2) a two-stage process in which the new MC portion was equated (using chained equating) to the reference MC portion using a MC-only anchor, and then the total composite scores were directly scaled in a single group design to the MC scores. In practice, the reference MC score was scaled to the reference composite score, whereas the new composite score was scaled to the new MC score. In this way, the linked new composite score would be on the scale of the reference composite score (see Figure 1C). The 417 examinees served as the reference population, and the 3,126 examinees served as the new form population. The lower panels of Figure 2 present the schematics of the two linking strategies.

---

2 Some may argue that there should be only one criterion, linear or nonlinear. This viewpoint appears reasonable. However, the linear criterion may also be viewed as the linear part in an expansion of the equipercentile function (von Davier, Holland & Thayer, 2004). If the true criterion function is indeed linear, then both the linear and equipercentile methods will yield the same answer. If the true criterion function is nonlinear, then the linear methods will be estimating the linear part of the equipercentile function, which is correctly represented by the linear criterion.
Evaluation. Both chained linear and chained equipercentile methods were used. 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 scores obtained using each linking method under each strategy were

---

**FIGURE 2. Schematic of the criterion and linking designs examined in this study.**

<table>
<thead>
<tr>
<th>Linking Strategy</th>
<th>Reference Group ((N = 417))</th>
<th>New Group ((N = 3,120))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion</strong></td>
<td><img src="image" alt="Ref Form" /> - <img src="image" alt="New Form" /></td>
<td></td>
</tr>
<tr>
<td><strong>Strategy 1:</strong></td>
<td><img src="image" alt="Ref Form" /> - <img src="image" alt="Anchor" /> - <img src="image" alt="Anchor" /> - <img src="image" alt="New Form" /></td>
<td></td>
</tr>
<tr>
<td>Direct Anchor Linking</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Strategy 2:</strong></td>
<td><img src="image" alt="Ref Form" /> - <img src="image" alt="Anchor" /> - <img src="image" alt="Anchor" /> - <img src="image" alt="New Form" /></td>
<td></td>
</tr>
<tr>
<td>Two-Stage Linking</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
compared with the criterion. The differences among the conversions were quantified using the weighted Root Mean Squared Difference (RMSD),

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

(3)

where \( i \) represents a raw score point, \( \hat{e}_i(x_i) \) is the equated scores of an equating method in a design at raw score \( x \), \( e_i(x_i) \) 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, a total of 500 bootstrap samples (i.e., 500 replications) were obtained using a resampling technique (e.g., SAS PROC SURVEYSELECT procedure). In each replication, examinees were randomly drawn with replacement from each reference and new form group until bootstrap samples consisted of the exactly same numbers of examinees as in the actual reference (\( N = 417 \)) and new (\( N = 3,126 \)) form groups. Then the new form scores were equated to the reference form for those 500 samples using both the direct and two-stage strategies with the chained linear and chained equipercentile methods.

In this case, equating bias was defined as the mean difference between chained equating and the criterion equating over 500 replications.

\[ Bias_j = \frac{\sum_{j=1}^{J} \left[ \hat{e}_j(x_i) - e(x_i) \right]}{J}, \]  

(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 the chained equating method in sample \( j \), and \( d_i \) is the
difference between \( \hat{e}_j(x_i) \) and \( e(x_i) \). 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.

\[
SEE_i = s(d_i) = \sqrt{Var[\hat{e}_j(x_i) - e(x_i)]} = \sqrt{Var[\hat{e}_j(x_i)]}.
\] (5)

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 Root Mean Squared Error (RMSE) index.

\[
RMSE_i = \sqrt{\hat{d}_i^2 + s(d_i)^2}.
\] (6)

As overall summary measures, we computed the weighted average root mean squared bias, the weighted average standard error of equating, and the weighted average RMSE across the new form group score distribution.

**Results**

Table 2 lists descriptive statistics for the composite tests, MC portions, and MC anchors in the new and reference groups. The table shows that the reference group outperformed the new form group on average on the MC anchor. The standardized mean difference was .16. The performance of the reference group on the new and reference forms (recall that this group had scores on both) indicates that the new form was somewhat easier than the reference form. The standardized mean difference between these two scores was .07. Table 2 gives the correlations between the MC anchor and other parts of the test for the new and reference forms. As expected, the MC anchor-total MC correlation was higher than the MC anchor-composite correlation. The correlations were fairly comparable across forms, giving an initial indication that (if the assumptions
of the research were correct) the tests might be successfully linked using MC-only anchors.

Table 2

*Means (Standard Deviations) for Examinee Groups Taking New and Reference Forms*

<table>
<thead>
<tr>
<th>Test score (Maximum value)</th>
<th>New Group (N = 3,126)</th>
<th>Ref Group (N = 417)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Form X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite (48)</td>
<td>31.3 (6.9)</td>
<td>33.7 (6.0)</td>
</tr>
<tr>
<td>Multiple Choice (16)</td>
<td>12.1 (2.4)</td>
<td>12.7 (2.2)</td>
</tr>
<tr>
<td>Old Form Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite (48)</td>
<td>—</td>
<td>33.3 (6.0)</td>
</tr>
<tr>
<td>Multiple Choice (16)</td>
<td>—</td>
<td>12.4 (2.1)</td>
</tr>
<tr>
<td>MC-only Anchor (8)</td>
<td>5.8 (1.3)</td>
<td>6.0 (1.3)</td>
</tr>
<tr>
<td>Composite-MC anchor correlation</td>
<td>.57</td>
<td>.55</td>
</tr>
<tr>
<td>Total MC-MC anchor correlation</td>
<td>.83</td>
<td>.83</td>
</tr>
</tbody>
</table>

Table 3 presents the difference between each linking function and the criterion, using the RMSD measure. According to this criterion, the two strategies (direct and two-stage) produced equivalent results. Figure 3 shows the results for direct anchor linking and for two-stage linking, plotted as conditional equated raw score differences from the criterion. The results for chained linear and chained equipercentile linking are shown on the same set of axes for convenience, although the criteria are different for those methods. The figure reflects that the direct and two-stage linking strategies yielded identical results in the linear case. In the nonlinear case, the differences between the two
strategies were negligible in the score range where most examinees were located. The figure also shows that in the nonlinear case, the two-stage strategy did not yield linked scores for the extremes of the score scale, although the direct method did.

Table 3

*Summary of Root Mean Squared Difference (RMSD) between Linking Results and Criterion for Two Linking Strategies with Linear and Equipercentile Methods*

<table>
<thead>
<tr>
<th>Method</th>
<th>Chained Linear</th>
<th>Chained Equipercentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linking Strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1: Direct (Chained Linking)</td>
<td>1.49</td>
<td>1.62</td>
</tr>
<tr>
<td>2: Two-Stage (Equating Plus Scaling)</td>
<td>1.49</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Table 4 summarizes the deviance measures – root mean squared bias, equating error, and root mean squared error – for the two linking strategies and two linking methods examined here. The table shows a substantial amount of bias for both linking strategies, on the order of one and a half score points. As one might expect, the equating error is slightly smaller for the linear than for the nonlinear linking models. There is nothing, however, to distinguish the direct anchor linking from the two-stage linking. This fact is illustrated as well in Figures 4 and 5, which show the conditional bias and CSEE for the linear and nonlinear linking procedures, respectively. In Figure 4, only one set of error bands is visible, because the results for the direct and two-stage methods lie directly on top of one another. For the nonlinear results shown in Figure 5, the direct and
two-stage results diverge in the lower part of the score scale, a region in which there are few people.

FIGURE 3. Difference between chained equating and the criterion for the direct and two-stage strategies, chained linear and chained equipercentile linking methods
### Table 4

*Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for the Two Linking Strategies and Two Methods*

<table>
<thead>
<tr>
<th>Linking Method/Strategy</th>
<th>Deviance measure</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>Equating Error</td>
<td>RMSE</td>
<td></td>
</tr>
<tr>
<td><strong>Chained Linear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct : Chained Linking</td>
<td>1.505</td>
<td>.432</td>
<td>1.566</td>
<td></td>
</tr>
<tr>
<td>Two-Stage: Equating and Scaling</td>
<td>1.505</td>
<td>.432</td>
<td>1.566</td>
<td></td>
</tr>
<tr>
<td><strong>Chained Equipercentile</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct: Chained Linking</td>
<td>1.602</td>
<td>.632</td>
<td>1.722</td>
<td></td>
</tr>
<tr>
<td>Two-Stage: Equating and Scaling</td>
<td>1.587</td>
<td>.609</td>
<td>1.700</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 4. Difference of linking results from the criterion for the direct and two-stage strategies, chained linear case.
The results of this study indicated substantial bias in equating. One of the suppositions underlying this study was that, even if the MC and the CR portions of the test did not measure the same construct, it might still be possible to use an MC-only anchor to link the reference and new composites, so long as the relationship between the MC portion and the composite were consistent across the reference and new form groups. The similarity of the MC-total correlations across the two groups gave an initial indication that the relationship might be consistent. To examine the consistency of the relationship across the two groups in more detail, the MC and composite scores were first placed on the reference form scale using score conversions obtained through direct
scaling in the reference population (recall that this group had scores on all measures). Next, the composite score was regressed on the total MC score in each group separately using a cubic regression function. Figure 6 displays the results.

Figure 6 reveals a substantial difference in the predicted conditional total score means across the two groups. At the lower end of the MC score range, this difference is almost five score points. Across most of the score range, the difference is greater than one score point. This difference in the MC-composite score relationship across the new and reference groups could easily result in the substantial linking bias seen here.

FIGURE 6. Cubic regression functions relating composite score to total multiple-choice score for the reference and new groups.

Note: Scores are placed on the scales of the reference forms.
Discussion

This study examined one possible solution to the problem of equating tests containing both MC and CR items, that of using a MC-only anchor. The study examined two strategies for accomplishing this. One involved using the MC anchor to link composite scores across the new and reference forms. The other used the MC anchor to link the new and reference form MC portions, to which the composite scores for both forms were then linked. The research showed that the two linking strategies yielded equivalent results when chained methods were used.

It may be that the direct and two-stage strategies would yield divergent results if post-stratification methods were used. In the present study, the very short MC anchor displayed relatively low correlation with the composite score. However, the correlation with the total MC score was much more substantial, as evidenced in Table 2. Post-stratification methods, such as Tucker linear equating, which make use of the bivariate moment in computations, would yield more satisfactory results when linking the MC scores as compared with the composite scores by virtue of the stronger correlation. Thus, we might expect an advantage for a two-stage process in this instance. Future research should investigate this possibility.

Linking performed on the data set in this study yielded quite biased results. In this sense, the results agree with those of previous researchers (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999). It could be that the linking methods will only be successful if the anchor and the composite score measure exactly the same constructs. Alternatively, it could be, as asserted in this paper, that the composite scores can be successfully linked even if the anchor does not measure exactly the same construct as the composite, in a
situation where the relationship between the anchor and the total composite score remains constant across groups.

The similarity of the MC-total correlations gave an initial indication that the relationship was constant across groups. However, the cubic regression results indicated that the MC-composite relationship varied across the new and reference groups. Nonetheless, the results here are equivocal; one cannot conclude from this evidence alone that invariance of the MC-composite relationship across groups (even in the presence of multidimensionality) will allow successful linking. A simulation study examining this possibility would shed more light on the issue.

An interesting finding emerged in the nonlinear case. When the chained equipercentile method was applied directly to the composite scores, the resulting new-to-reference form score conversion table covered all possible raw score points on the new form. When the two-stage strategy was employed, however, the score scale was truncated at the top and the bottom, such that no converted scores resulted for the extremes of the new form score scale. This truncation is most likely a result of the concatenation of the linking and scaling results, combined with the sparseness of the data in the extremes of the distributions. Further investigation is warranted to determine the root cause of the phenomenon. However, given the similarity of the results from the two linking strategies, it might appear most prudent to apply the direct anchor method as opposed to the two-stage method.

A practitioner may still prefer the two-stage method for logistic reasons.
Whenever a test contains CR items, test linking is delayed while CR items are scored. Using a two-stage strategy would allow for partial linking of test scores before the CR
item scores are available. Especially when multiple forms must be equated in a restricted
time frame, a two-stage strategy could allow for expedited production of test score
conversions once CR scores become available.
References


