Comparisons of Test Characteristic Curve Alignment Criteria of the Anchor Set and the Total Test: Maintaining Test Scale and Impacts on Student Performance

Thakur B. Karkee, Ph. D.
Measurement Incorporated

Kevin Fatica
CTB/McGraw-Hill

Stephen T. Murphy, Ph. D.
Pearson

Paper presented at the annual meeting of the
National Council on Measurement in Education (NCME)
Denver, Colorado
April 29 – May 4, 2010

1. Introduction

In the context of the *No Child Left Behind Act* (2002) and associated Adequate Yearly Progress (AYP) requirements, many state testing programs now place considerable emphasis on the ability to compare test scores across years. Theoretically, one could use the same test form in multiple administrations but that leads to testing concerns, especially item exposure, and the related question of whether the obtained score is measuring real student ability or is a product of advance knowledge of the items on the test. In order to avoid such concerns, most states create equivalent test forms for use across years, which enable them to produce comparable scores from year to year.

Among states that use item response theory (IRT) in their testing programs, the most common approach to developing equivalent forms for use across years is the *common item non-equivalent group design.* In this design, the test from the baseline year and a subsequently created new test form contain some common items. The common items are also often referred to as anchor items. Using a process called equating, the set of anchor items can be used to place the new test form onto the baseline scale, thereby making the test scales, and the test scores on those scales, equivalent (See Kolen & Brennen, 2004; Stocking and Lord, 1983). Setting aside measurement error, equivalent test forms produce equivalent scores. When test forms have been successfully equated, scores from those tests can be used interchangeably.

Anchor sets used in equating are subjected to considerable scrutiny, both among test developers, and as noted in the psychometric literature. There are a number of guidelines available which specify appropriate characteristics for anchor sets. For example, the Council of Chief State School Officers (CCSSO, 2003) provides a list of guidelines specifying that the anchor items should be representative of the total test. Specifically, CCSSO recommends that the anchor set should represent the test blueprint, (i.e., be a miniature version of the total test in terms of content and item characteristics); be free from bias and poor fit; and in summary, be the best items in the item pool.

However, recent research suggests that we should reconsider some of these traditional guidelines. For example, while it has been a generally held understanding that the spread of item difficulties in the anchor set should mirror the spread of item difficulty in the total test, a recent work by Sinhary and Holland (2007) suggests that this view should be reconsidered. Through a series of simulations as well as real data, and using unidimensional and multidimensional IRT models, Sinhary and Holland demonstrated that when the level of item difficulty in the anchor set spreads across the range of item difficulty in the total test, the raw score correlation between the anchor set and the total test was consistently *lower* than when the item difficulty was in a narrower range. As such, Sinhary and Holland (2007) suggested that using an anchor set that mirrors the range of difficulty in the total test was not an optimal guideline for developing an anchor set.

The current paper investigates a tenet of the traditional view on the psychometric characteristics of such anchor sets. Specifically, the traditional guideline, without any specificity, states that the test characteristic curve (TCC) of the anchor set and the total test should be closely overlapped. A general rule of thumb regarding the overlapping TCCs is that, for any given scale

score, the expected proportion of the maximum raw score (EPMRS) difference based on the TCC of the anchor set and the TCC of the new test form should be 5% or less. Theoretically, the TCC models the relationship between an ability level, or theta level, and a raw score on the test. For every level of the ability, the TCC identifies the expected proportion of the raw score to be obtained on the test. The 5% EMPRS difference criterion implies that if a new test has100 raw score points then a difference of five raw score points between the TCCs produced from the anchor set and the new test across the ability continuum would be allowed.

Note that the 5% EPMRS difference criterion is arbitrary and it can be met in a number of ways. For example, the TCCs of the anchor set and the total test could have exactly the same shape with a 5% difference at all points along the ability continuum. Alternatively, the two TCCs might match very well at the low end of the ability continuum while at the high end the difference expands to 5%. The opposite scenario could also occur: the TCCs could show larger differences all along the lower end of the ability continuum, but then tighten up in the middle of the ability range and fade away to near zero at the high end of the ability continuum. It could also be that the TCCs align well in the middle of the ability continuum, but at both the high end and the low end of the ability range the difference expands to 5%. Clearly, many possibilities exist. And yet, other than this general rule of thumb, the psychometric literature offers little formal guidance in terms of how closely the TCCs of the anchor set and the total test should be aligned.

The current study engages in an exploratory evaluation of this 5% EPMRS difference criterion. In order to explore and evaluate the utility of using this 5% criterion, the current study constructs a series of anchor sets that *meet, exceed,* and *violate* this criterion, and then these anchor sets are used to scale and equate a single test form from a recent large scale assessment in a year-to-year common item non-equivalent groups equating design. The differential impacts of using these different anchor sets are evaluated in order to consider the relative impact of meeting, exceeding, and violating the 5% criterion.

In the current study, the impacts of using the different anchor sets are evaluated on two fronts: first, we consider the characteristics of, and differential impacts of, the anchor sets from a psychometric perspective. We looked specifically at:

- anchor set TCCs before and after equating;
- patterns of scaling constants (M1, M2); and
- correlation coefficients of item parameters: a (discrimination), b (location), c (pseudo-guessing), and p (probability of answering the item correctly) parameters before and after equating.

Second, we considered what state departments of education are centrally focused on: impacts on student scores and proficiency classifications, and the stability or equivalence of student scores as estimated by different test forms. More specifically, for each anchor set we looked at:

- percentile ranks (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 99^{th}) of scale scores;
- percentile ranks at the cut scores;
- scale score distribution; and
- percentages of students classified in each proficiency level.

As described, the current study offers insight into how the degree of similarity between the anchor TCC and the total test TCC impacts test scales and student scores. This study also considers the utility of using this 5% criterion in test development. As such, the current study is of interest to both test developers and users of tests, specifically the large scale testing programs that use the common items non-equivalent groups equating design. Test development is a challenging enterprise that navigates a difficult path among a number of criteria, guidelines, and considerations, all while simultaneously facing the constraints of budgets and the limitations of a given item pool. For states, test developers, and other users of tests like these, the possibility that one of the more difficult criteria of test development could warrant reconsideration presents an opportunity not just to improve practice from a theoretical and psychometric point of view, but from the point of view of reducing the costs for the test development

2. Methods

2.1 Data

The data for this study came from a recent large-scale assessment for grade 8 mathematics. Over 50,000 students participated in the assessment. The assessment consisted of 60 items with approximately 75% multiple-choice (MC) and 25% constructed-response (CR) items. The test consisted of 20 MC anchor items that were internal to the test, meaning they contributed toward the student's score on the test.

2.2 Anchor Set Selection

Five anchor sets were selected from the available item pool. We developed anchor sets that met the 5% EPMRS difference criterion exactly, anchor sets that did better than the 5% difference, and anchor sets that violated this 5% criterion. More specifically, we developed anchor sets where the EPMRS difference was 2%, 4%, 5%, 6%, and 8%. We refer to these anchor sets as sets S1, S2, S3, S4, and S5 respectively (Table 1).

Table 1. Anchor Sets and Corresponding EPMRS Differences

Anchor Set	EPMRS Difference
S1	2%
S2	4%
S3	5%
S4	6%
S5	8%

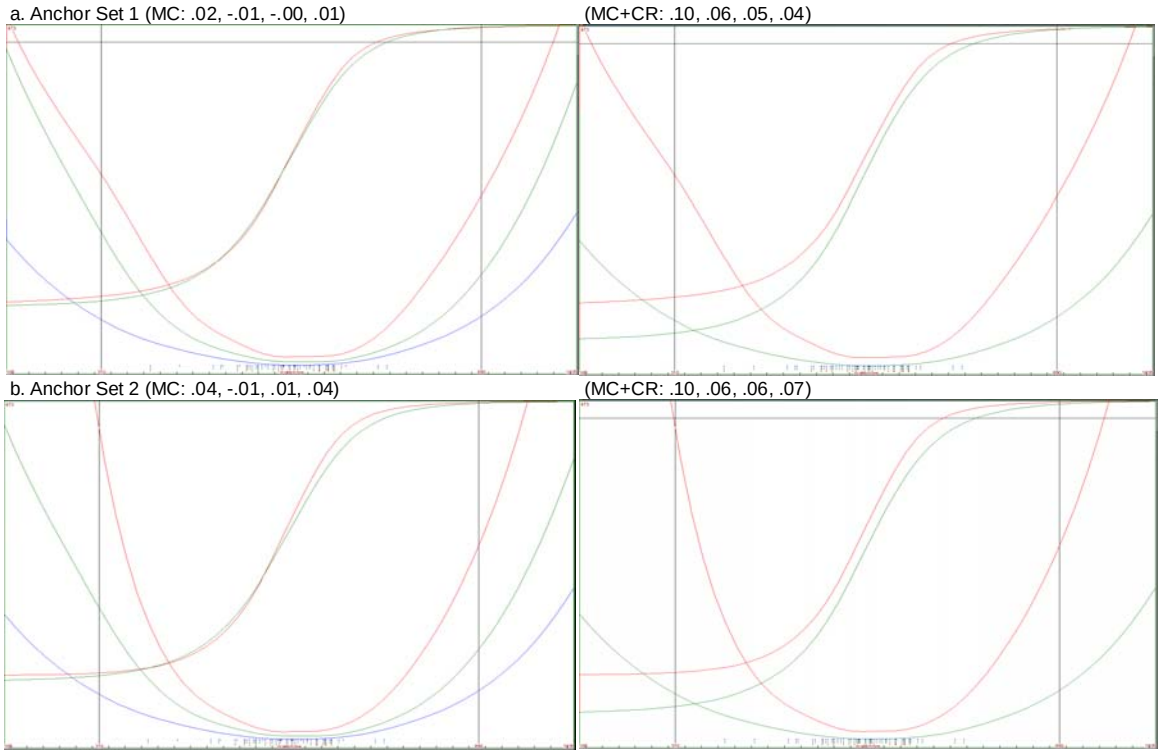
All of the anchor items were MC, and all of the anchor sets contained 20 items each. To create the anchor sets, 20 items were selected first to create S1, where the EPMRS difference was 2%, and then the remaining four anchor sets were created by replacing a portion of the items

in S1 with items from the available item pool. The replacement was made in such a way that the content representation was maintained as closely as possible. Each anchor set therefore, contained some items in common. The anchor items were all operational items, meaning that they were part of the newly developed test for the succeeding administration.

2.2.1 Test Characteristics Curves Comparison

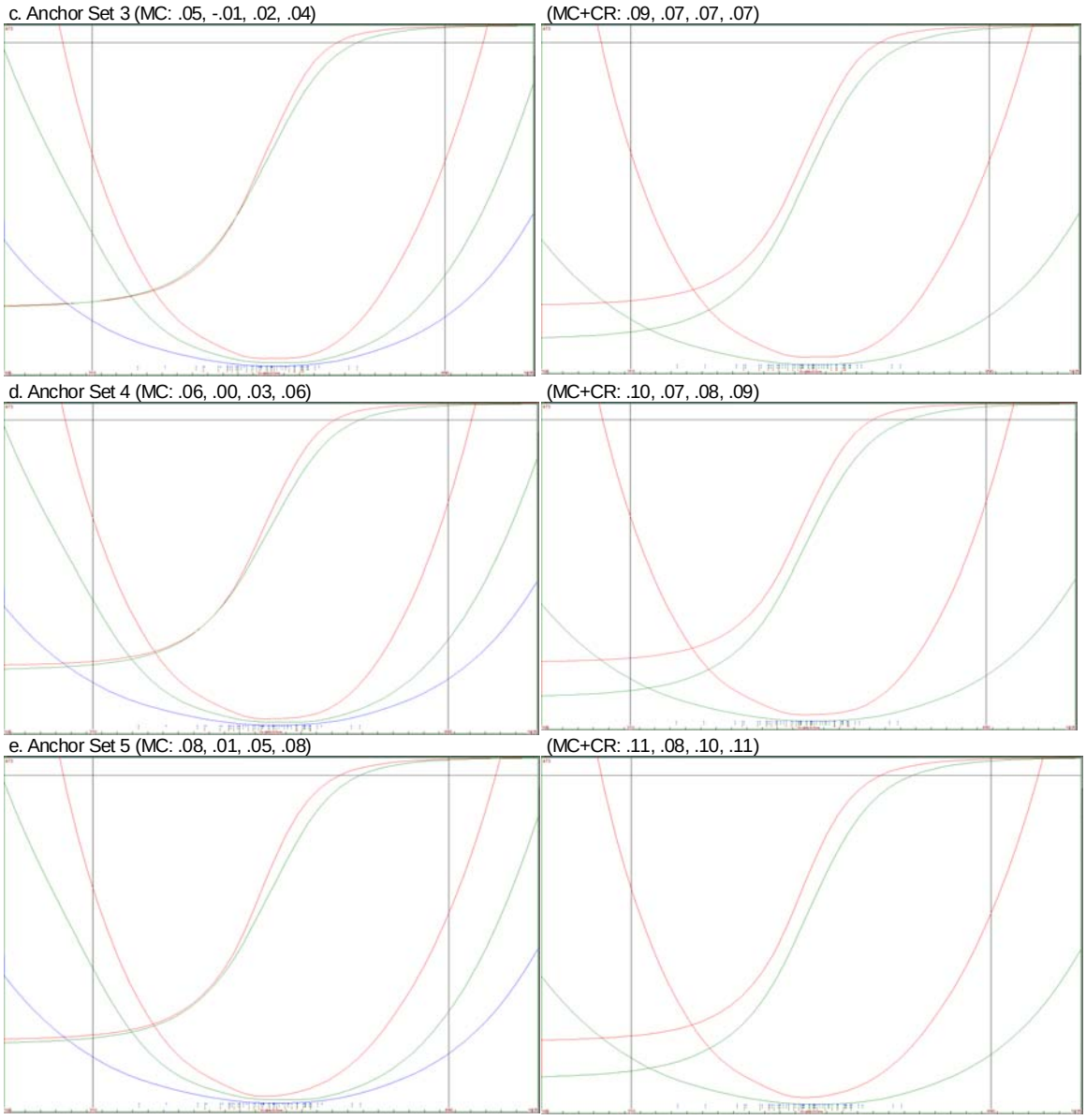
The TCCs of each anchor set and the total test are presented in Figure 1(1a-1e). The EPMRS differences between the anchor set and the total test were based on the MC items only, as is consistent with current practice for the assessment. TCCs based on MC items only are presented on the left. TCCs where the total test includes both MC and CR items are presented on the right for reference purposes. The header line for each figure shows the overall EPMRS difference, followed by the EPMRS difference at the three cuts. For example, in Figure 1a the overall EPMRS difference is .02, the difference at the Basic cut is -.01, the difference at the Proficient cut is -.00 and the difference at the Advanced cut is .01. Negative differences indicate that the EPMRS obtained for a given scale score is higher in the anchor set than in the reference test. Positive differences indicate that the EPMRS obtained is lower in the anchor set than in the reference test. The U shaped curves are, from narrowest to widest, the standard error (SE) curves of the anchor set, the total test, and the SE of all items in the item pool.

Figures 1a to 1e.
Anchor Set (left) and Total Test (right) TCCs



Note: The U shape curves are standard error (SE) curves associated with anchor set, total test, and item pool

Figures 1a to 1e. (Continued)
Anchor Set (left) and Total Test (right) TCCs



Note: The U shape curves are standard error (SE) curves associated with anchor set, total test, and item pool

Figure 1 above shows that in virtually all cases for all anchor sets the EPMRS differences were positive, except at the Basic cut in anchor sets S1, S2, and S3. This indicates that the anchor sets were, with few exceptions, consistently easier than the total test at all points on the ability continuum. With reference to the TCCs based on MC items only (the left side of Figure 1), Figure 1 also shows that for each anchor design, the TCCs of the anchor set and the total test were very close in the low to middle ability range. The differences between the two TCCs tended to be on the higher end of the ability range. When we look at the right side of the figure, where TCCs for the total test include both MC and CR items, we can see that differences in TCCs were

larger across the ability continuum with relatively smaller differences in the middle of the ability range.

The EPMRS differences in the TCCs are also tabulated in Table 2. Note that overall EPMRS difference and differences at the cuts are smallest for anchor set S1. The differences between the anchor set and the total test were intentionally increased from anchor set S1 to anchor set S5, (as noted above, the calculation of that EPMRS difference was based on MC items only in the total test). Therefore, at the cut score for the Advanced performance level, the difference in EPMRS moves from 0.01 in S1, to 0.08 in S5. Figures 1a-1e and Table 2 also show that when comparing the anchor set to the total test based on both MC and CR items, the overall 5% difference criterion was violated in all anchor sets and at every cut score level in each anchor set, except in S1.

Table 2. Difference of EPMRS Between the Two TCCs

Anchor Set	Test	Overall	Basic	Proficient	Advanced
S1	MC	0.02	-0.01	0.00	0.01
	MC+CR	0.10	0.06	0.05	0.04
S2	MC	0.04	-0.01	0.01	0.04
	MC+CR	0.10	0.06	0.06	0.07
S3	MC	0.05	-0.01	0.02	0.04
	MC+CR	0.09	0.07	0.07	0.07
S4	MC	0.06	0.00	0.03	0.06
	MC+CR	0.10	0.07	0.08	0.09
S5	MC	0.08	0.01	0.05	0.08
	MC+CR	0.11	0.08	0.10	0.11

Note: MC=Multiple-Choice; CR=Constructed-Response

2.2.2. Blueprint Representation

Each anchor set was also constructed to maintain the same level of content representation. The content representation per content standard based on the total test and each anchor set is presented in Table 3. Note in Table 3 that the anchor sets mirror the content representation of the total test. The anchor set S3, however, under represents Standard 1 by 10%, and over represents Standard 2 by 10%. Similarly, the anchor set S5 also over represents Standard 2 by 10%. These changes in content representation resulted in from limitations in the available item pool.

Table 3. Test Blueprint Represented by the Anchor Sets

				Anchor Set % (N=20)				
Standards	Total Items	Total Pts.	% of Total Pts.	S1	S2	S3	S4	S5
1	15	22	25.3%	20%	20%	15%	20%	20%
2	15	22	25.3%	25%	30%	35%	30%	35%
3	12	17	19.5%	20%	20%	20%	15%	20%
4	18	26	29.9%	35%	30%	30%	35%	25%

2.3 Calibration and Equating

The five anchor sets were created using the item parameters from a reference test. The MC item and CR item response data on the new form were calibrated simultaneously using the three-parameter logistic (3PL) model (Lord, 1980; Lord & Novick, 1968) for MC items and the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) for the CR items. The item parameters in the new form, in the theta metric, were linked to the reference scale in the scale score metric using the five anchor sets.

The Stocking and Lord (1983) procedure was used for the linking. The Stocking and Lord (SL) procedure, also known as the Test Characteristic Curve method, determines the scaling constants (the multiplicative constant, M1 and additive constant, M2) in such a way as to minimize the average squared difference between the true score estimates. That is, M1 and M2 based on the SL procedure can be found by minimizing the quadratic loss function (F):

$$F = \frac{1}{N}\sum_{a=1}^{N}(\hat{\psi}_j - \hat{\psi}_j^*)^2$$

Where, $\hat{\psi}_j$ is the estimated true score obtained from the 3PL model and $\hat{\psi}_j^*$ is the estimated true score obtained from the 3PL model after it has been transformed to the previous scale as follows:

$$\hat{\psi}_j = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; a_i, b_i, c_i)$$

$$\hat{\psi}_j^* = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i)$$

This method has been used widely in large-scale assessments and has been demonstrated to be similar to or better than the mean/mean, mean/sigma, and Haebara methods (Karkee & Wright, 2004; Baker & Al-Karni, 1991; Hung et al., 1991; Way & Tang, 1991).

2.4 Results Evaluation Criteria

The transformed item parameters for the new form, based on the five designs, were used to score student responses using the item pattern scoring method to compute student scores in the test. After the linking was completed, the results across the five anchor designs were evaluated in terms of:

- Existence of any systematic patterns in the scaling constants (M1, M2).
- Correlation coefficients of the following item parameters before and after equating: a (discrimination), b (location), c (pseudo-guessing), and p (proportion correct).
- Any significant change in the anchor set TCCs before and after equating.
- The scale score distribution and root mean square deviation (RMSD) of scale scores.
- The percentile ranks of scale scores at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 99^{th} percentiles, and the percentile ranks at the cut scores. Note: we would expect to see differences in the percentile rank of the scale scores if the differences in the anchor sets have impacted the test scale.
- Differences in the percentage of students classified into different proficiency levels. For each design, the same cut scores were applied to classify students into one of the four proficiency categories, and then the percentage of students in each proficiency level was compared.

3. Results

3.1 *Scaling Constants and Parameters Correlation Coefficients*

As described above, the SL procedure determines the scaling constants M1 and M2 in such a way that the average squared difference between the true score estimates based on the anchor set and the total test is as small as possible. The scaling constants produced are shown in Table 4. Table 4 also shows the correlation coefficients between the anchor item parameters in the total test before and after equating.

Table 4. Scaling Constants and Parameter Correlation Coefficients

Anchor Set	M1	M2	Correlations			
			a	b	c	p
S1 (Diff.=.02)	61.95	276.62	0.98	0.92	0.65	0.99
S2 (Diff.=.04)	62.08	278.36	0.98	0.87	0.78	0.99
S3 (Diff.=.05)	62.33	278.52	0.99	0.99	0.98	0.99
S4 (Diff.=.06)	61.67	277.04	0.98	0.90	0.78	0.99
S5 (Diff.=.08)	61.64	277.87	0.98	0.88	0.80	0.99

Note: M1=Intercept; M2=Slope; a=discrimination; b=location; c=pseudo-guessing factor; p= proportion correct.

Results indicated that the scaling constants across the anchor designs were very similar. Recall that the equating procedure adjusts the difficulty of the two tests (Kolen & Brennan, 2004). If there was a systematic relationship between the EPMRS differences and the difficulty of the tests, given that the anchor sets were consistently easier than the total tests, we would expect to see progressively easier tests as the EPMRS difference grows larger. However, while the tests based on the anchor sets S4 and S5 were easier (based on M2) than the tests based on anchor sets S2 and S3, the easiest test is actually the test that uses anchor set S1, where the pull of the anchor set toward an easier test was, theoretically, the weakest overall. Overall, the results, in terms of scaling constants, do not show a consistent pattern or systematic relationship with the EPMRS difference in the anchor set and the total test TCCs.

Similarly, most *a, b, c,* and *p* parameters correlation coefficients were reasonably high, and there was no substantial, consistent, or systematic variation in the correlation of the *a*, *b*, *c*, and *p* parameters corresponding to the systematically increasing EPMRS differences created in the anchor set and the total test TCCs.

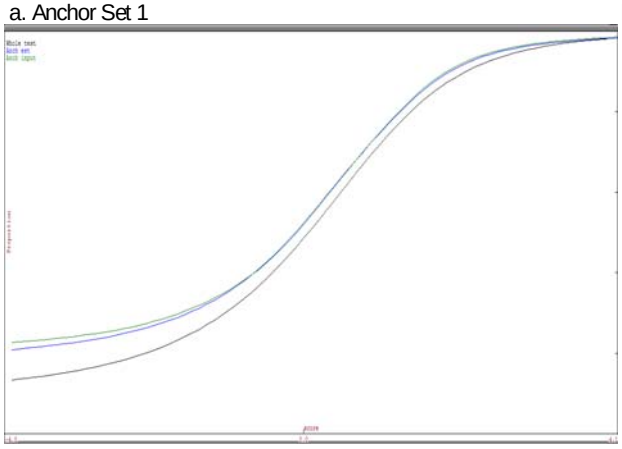
3.2 Anchor Set TCCs Before and After Equating

Figure 2 (2a to 2e) presents the TCCs of the anchor sets before and after equating and the TCCs for the new total tests with MC and CR items. These figures show that the TCCs for the anchor items before and after equating overlapped closely; although for each anchor set, at the lower ability range, the TCCs before and after equating showed some separation, albeit minimal. These results indicate that the anchor item parameters were reasonably stable between the two

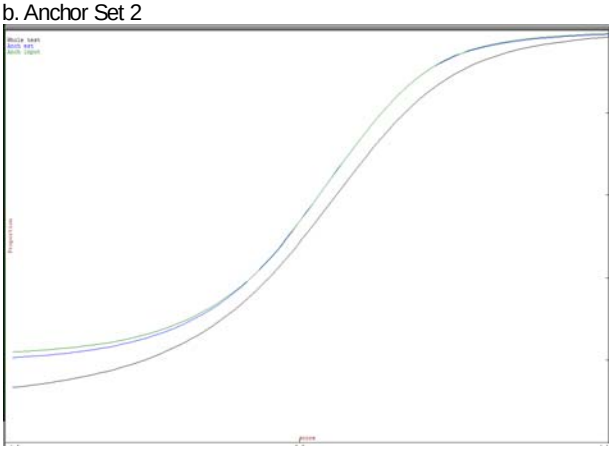
administrations. Also, as depicted in Figures 1a to 1e the anchor item TCCs are located to the left of (easier than) the new total test for each anchor set.

Figure 2. TCCs of Anchor Sets Before and After Equating and Total Test (MC +CR Items)

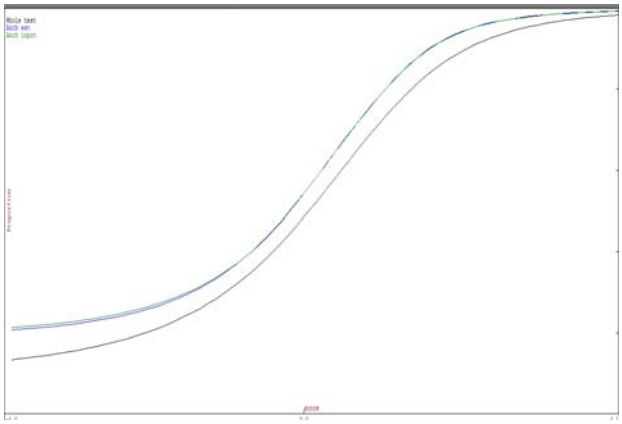
a. Anchor Set 1



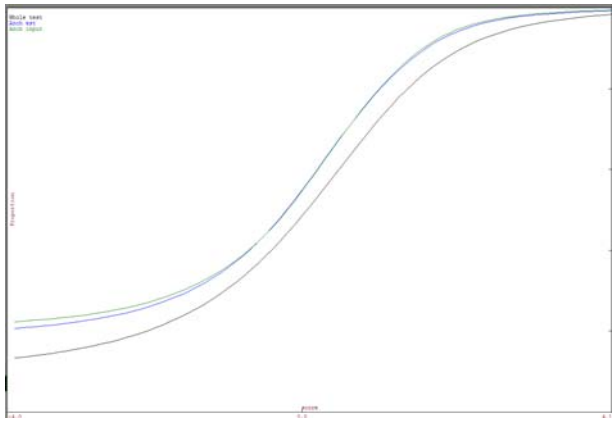
b. Anchor Set 2



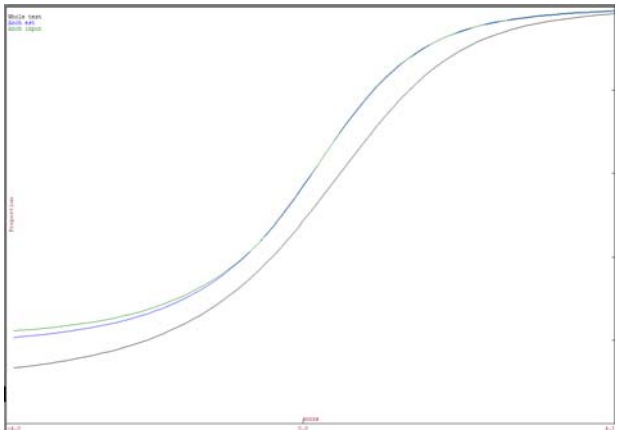
c. Anchor Set 3



d. Anchor Set 4



e. Anchor Set 5



3.3 Scale Scores Descriptive Statistics

For each anchor design, the responses for approximately 50,000 students in the total test were scored using the transformed item parameters, after linking to the common scale. The mean and standard deviation of the scale scores across the five anchor designs are shown in Table 5. If there was a systematic relationship between the EPMRS difference and the mean scale score, given that the anchor sets were consistently easier than the total tests, we would expect to see a progressively higher mean scale score as the EPMRS difference grows larger and as the anchor set, theoretically, pulls the total test further towards an easier test and a higher mean scale score. However, the observed scale score means are very similar across the designs, and the mean scale score results do not reflect any substantial, consistent, or systematic variation corresponding to the systematically increasing EPMRS differences between the anchor set and total test TCCs.

Table 5. Mean and Standard Deviation of Scale Scores

Sample	Mean	Std Dev
S1	270.18	70.94
S2	271.90	71.10
S3	272.04	71.37
S4	270.62	70.65
S5	271.47	70.64

3.4 *Root Mean Square Deviation*

A two-way scale score root mean square deviations (RMSD) table is presented in Table 6. The RMSD here measures the average scale score differences among the five anchor designs. The values in each cell indicate the RMSD of scale scores between the corresponding designs. As was observed in other results, there was no particular trend corresponding to the systematically increasing EPMRS differences created in the anchor sets, nor was there any trend indicating that one anchor set was consistently different from the other anchor sets.

Table 6. RMSD of Scale Scores Between the Anchor Designs

	S1	S2	S3	S4
S2	1.79			
S3	1.96	0.50		
S4	0.69	1.42	1.65	
S5	1.41	0.76	1.02	0.92

3.5 Rank Order of Scale Scores and Percentile Rank at Cuts

Table 7 shows the rank order of scale scores and percentile ranks at the cut scores. If there was a systematic relationship between the EPMRS difference and the scale score at the cuts, given that the anchor sets were consistently easier than the total tests, we would expect to see progressively higher scale scores at the cuts as the EPMRS difference grows larger. Looking across the anchor designs, one can observe that the scale scores for the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 99^{th} percentiles in anchor sets S2, S3, S4, and S5 are equal to or larger than the scale scores in S1. This seems consistent with a modest or weak relationship between the EPMRS difference and the scale scores at the cuts. However, since these differences are too small and inconsistent, they do not reflect any systematic variation corresponding to the systematically increasing EPMRS differences created in the anchor set and the total test TCCs.

Similarly, Table 7 further shows that the percentile rank at the cut scores remained the same or decreased for the anchor designs with larger EPMRS differences. These differences are also slight and inconsistent, and as with the other results noted above, they are not interpreted here as reflecting any substantial, consistent, or systematic variation corresponding to the variation in EPMRS created in the anchor sets. The maximum rank order difference was only one (1) percentile point.

Table 7. Rank Order and Percentile Ranks at Cuts

Rank Order	S1 (Diff.=.02)	S2 (Diff.=.04)	S3 (Diff.=.05)	S4 (Diff.=.06)	S5 (Diff.=.08)
10	183	184	184	184	185
25	229	230	230	229	230
50	273	275	275	273	274
75	317	319	320	318	319
99	420	423	424	421	422

Percentile Rank at Cut					
	Percentile Ranks for Different Anchor Designs				
Cuts	S1	S2	S3	S4	S5
221	22	21	21	22	21
277	53	52	52	52	52
328	80	79	79	80	80

3.6 Impact of Different Anchor Designs on Proficiency classifications

Table 8 shows the percentage of students classified into different proficiency levels, based on the approximately 50,000 examinees. For the anchor set S1 (2% EPMRS difference), 21.53% were classified into the Minimal proficiency level, 30.63% were Basic, 27.42% were Proficient, and 20.42% were in the Advanced proficiency level. The proficiency classifications were very similar across all five anchor designs. There was a maximum difference of 0.64% at the Minimal level, 0.53% at the Basic level, 0.32% at the Proficient level, and 1.01% at the Advanced level. Note that the maximum difference was larger for the Advanced category, but

the next largest difference was in the Basic category. It is worth reminding readers here that the increasing EPMRS differences in the anchor sets and the total test TCCs were created mostly by pushing the anchor set TCCs away from the total test TCCs at the higher ability range. Were we to see a progressively higher proportion of students in the Advanced level, that could be taken as suggestive of some relationship between EPMRS differences in anchor sets and the total test TCCs. However, the evidence is inconsistent. Overall, the different anchor designs showed little impact on the classifications of students into the different proficiency levels.

Table 8. Impact of Anchor Designs on Proficiency Classifications

	Minimal	Basic	Proficient	Advanced	At or Above Proficient
S1 (Diff.=.02)	21.53	30.63	27.42	20.42	47.84
S2 (Diff.=.04)	20.89	30.22	27.61	21.28	48.89
S3 (Diff.=.05)	20.91	30.10	27.55	21.43	48.98
S4 (Diff.=.06)	21.29	30.59	27.64	20.48	48.12
S5 (Diff.=.08)	20.91	30.47	27.74	20.87	48.61

Note: all values are in percentages

4. Summary and Discussions

In large-scale assessments where student scores are used for high-stakes decision making, maintaining the scale established in the baseline year is crucial for year to year score comparability.

The current study investigated the traditional requirement that the EPMRS difference in the TCC of the anchor set and the total test be 5% or less. Using a series of anchor sets that a) met the 5% criterion exactly, b) did better than the 5% criterion, and c) violated the 5% criterion, this study demonstrated that whether we met the criterion exactly, exceeded it, or violated it, the ultimate results varied little. As such, the main conclusion of this study is that the widely held 5% EPMRS difference criterion warrants reconsideration and further study.

Among the main findings of this study were that, in terms of the scaling constants, the TCCs of the anchor sets before and after equating, and the anchor item parameter correlation coefficients, offered no clear evidence that the violation of the 5% EPMRS difference criterion would produce results significantly different than anchor designs that met or exceeded the criterion. While the additive constants based on the anchor sets with the larger EPMRS differences were larger than those from the anchor design with smallest difference, there was no consistent pattern suggesting that larger differences in the EPMRS bear a systematic relationship with larger additive constants.

The mean scale scores and the corresponding percentile ranks also did not reflect any substantial, consistent, or systematic variation corresponding to the EPMRS differences created in the anchor set and total test TCCs. The same was true for the percentile rank order at the cuts.

And, as noted, the percentage of students classified into different proficiency levels was comparable across all anchor designs.

In terms of blueprint representation, the anchor designs S3 and S5 departed by about 10% for some standards, but the results did not show any indication that the violation impacted student scores. In addition, the correlation coefficients of the item parameters before and after equating were reasonably high and comparable across all anchor designs.

In conclusion, the results indicated that, whether the EPMRS difference *met, exceeded, or violated* the 5% criterion, the results were very similar across the five anchor designs. As such, the main conclusion of this study, as indicated above, is that this 5% EPMRS difference criterion warrants reconsideration and further study. The conclusions of this study should be of interest to test developers and users of tests, specifically the large scale testing programs that use the common items non-equivalent groups equating design.

In closing, having established that the generally held view on the usage of the 5% EPMRS difference criterion warrants reconsideration, we offer a few possible directions along which this line of inquiry could be further pursued. First, future research could consider how the results provided in the current study might be different were the number of items in the anchor set to vary. For example, we could ask how much the size of the anchor set relative to the size of the total test may either mitigate or enhance any influence of the anchor set on the total test TCC. We could ask, in that context, how any influence of the anchor set TCC would change if, rather than being about one-third the size of the total test, the anchor set were only 15% the size of the total test, or even 10%? And if the influence of the anchor set were weaker where the anchor size is smaller relative to the total test, would this 5% EPMRS difference criterion become more crucial in that context?

Second, the authors suggest that the inquiry raised here might be advantageously pursued through simulations. The current study was based on real data, and while that provides certain advantages, this arrangement also limits some of the conditions that might be explored more easily in controlled simulation studies. For example, it would be interesting to consider the implications of a scenario where the anchor set TCC is more difficult than the total test TCC, but only in one or two standards. Simulation studies may also enable researchers to examine some extreme cases of EPMRS differences, and thereby gain additional insight into the ways these differences in EPMRS may influence total test TCCs in practice. Clearly, there are a number of potentially interesting possibilities to consider here. Hopefully, this paper is just the beginning of what will be a fruitful line of research, which, in the future, includes additional controlled simulation studies that can further inform both theory and practice on this important issue.

References

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Council of Chief State School Officers (CCSSO, 2003). Quality control checklist for item development and test form construction. *Technical Issues in Large-Scale Assessment*.

Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking.* Paper presented at the International Academic Symposium on Psychological Measurement, Tainan, Taiwan.

Karkee, T. B. & Wright, K. R. (2004). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale*. Paper presented at the Annual Meeting of the American Educational Research Association in San Diego, California, April 16, 2004.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

No Child Left Behind Act (2002). No Child Left Behind Act of 2001, Pub. L. No. 107−110, 1 U.S.C. (2002).

Sinhary, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement Fall 2007*, 44(3), 249-275.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Way, W. D., & Tang, K. L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Yen, W. M. (1985). Increasing Item Complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item independence. *Journal of Educational Measurement*, *30*, 187–213.