

Running Head:

IMPACT OF DIFFERENT ANCHOR STABILITY METHODS

The Impact of Different Anchor Stability Methods
on Equating Results and Student Performance

Stephen Murphy, Ian Little, Meichu Fan,
Chow-Hong Lin, & Rob Kirkpatrick

Paper presented at the annual meeting of the National Council of Measurement in
Education, Denver, CO, May 2010.



*Using assessment
and research to
promote learning*

The Impact of Different Anchor Stability Methods on Equating Results and Student Performance

Introduction

Dramatic changes in item statistics can result in systematic errors in equating results for common-item non-equivalent groups designs (we will refer to this as CINEG; Kolen & Brennan, 2004; also referred to as non-equivalent anchor test designs, Holland & Dorans, 2006). In an item response theory context it is customary to evaluate the differences in item parameters for common items estimated in old and new forms, and evaluate how those changes affect equating outcomes. The process used in this evaluation is often referred to as an anchor stability check and is one in a series of judgments that must be made during equating. If the evaluation concludes that there are large differences in the estimates for a particular item that item may be dropped from the anchor set and the scale transformation constants re-estimated with a reduced set of anchors. In practice, anchors may be removed in either a list-wise or step-wise fashion until the remaining items in the anchor set are considered stable. The scaling constants from this stable anchor set are used to place items from the new form onto the scale of the old form.

It is important to note that this anchor stability check may not be able to identify why there is a change in the item parameter estimates. This is because there are several sources of instability that are confounded in the differences among item parameter estimates. These include random estimation error, samples that are too far apart in terms of ability, item position change effects, differences in how the anchor items are presented to students on the two forms, differences in the dimensionality of the anchor set

compared to the full length forms, changes in curricular emphasis, and others. In practice, when unstable items are identified practitioners may ask test developers to investigate some of these potential causes. We note that removing items with unstable parameter estimates may itself be a source of equating inaccuracy, especially if the items under investigation reflect construct-relevant population change (Miller & Fitzpatrick, 2009). For the purposes of this research, we assume that anchor instability is construct-irrelevant, even though in practice it may not be.

There are at least four commonly used approaches to examining these differences. One method is to produce a bivariate plot with the old parameter values on one axis and the new values on the other, then to make a judgment about possible outliers using an visually-established ‘line of best fit’ approach and a graphical display. Outliers are removed until parameter estimates appear to fall uniformly around the ‘line of best fit.’ For 2PL, 3PL, and polytomous models a separate plot may be produced for each parameter in the model. Practitioners may emphasize differences in a particular parameter, perhaps difficulty (\hat{b}), while deemphasizing differences in another parameter such as lower asymptote (\hat{c}). A limitation of this approach is that there are no universally accepted criteria to determine what amount of departure from the ‘line of best fit’ reflects an outlier. Another limitation is that for models with more than one parameter, each parameter influences the others in the item parameter estimation process. Evaluating the item parameter estimates in isolation from one another may lead to faulty conclusions.

A second method is to drop items that exceed a heuristically developed critical difference in the parameter estimates. This method, most often used with Rasch or 1PL models is typically called the displacement method. In the displacement method items

from the anchor set are dropped if the difference between old and new is greater than some pre-specified value. A key limitation of this approach is that the pre-specified value is arbitrary. In practice, we have seen many variations of this approach, ranging from simple single-step implementations to quite complex multi-step implementations.

A third approach is to use robust z-values, which are computed by taking the difference between the pre-equated value and the post-equated value and subtracting the median value of the differences and dividing this number by the interquartile range multiplied by 0.74 (SCDE, 2001):

$$Robust\ z_j = \frac{(adif_i - MDdif_j)}{(IQ_j \times 0.74)} \quad (4)$$

where $adif_i$ = the difference between pre and post-equated parameters (\hat{a} or \hat{b}) for item i , $MDdif_j$ = the median of the differences for anchor set j , and IQ_j = interquartile range for anchor set j . This method is commonly used for Rasch models, though it might be applied to 3PL models as well.

An alternate approach to examining anchor stability is what we refer to as the d^2 method that employs a procedure analogous to examining differential item functioning. For each anchor item, the weighted sum of the squared deviation between the Item Characteristic Curves (ICC; d^2) is calculated based on old and new parameter estimates at each point of a theoretical theta distribution:

$$d_i^2 = \sum^k [P_{ix}(\theta_k) - P_{iy}(\theta_k)] \cdot g(\theta_k) \quad (5)$$

where, i = item, x = old, y = new, k = theta quadrature point, and g = theoretically weighted posterior theta distribution. The ‘initial’ Raw Score to Scale Score (RSSS) table is computed for the anchor set with no removed items. The d^2 are heuristically reviewed to determine outliers. Presently, no standard criterion is used, though two approaches are common: a) flag as outlier items that exceed the 95th percentile of an anecdotal distribution of d^2 (typically d^2 values for a project over years of administration), and b) rank order the items by d^2 and flag as outlier the item with the greatest magnitude if the ratio between this item and next highest d^2 item exceeds some arbitrary value (say 1.5). If there are no outlier d^2 values, no items are considered unstable. Otherwise, items are removed from the anchor set in either a list-wise or step-wise fashion. One step-wise fashion that we use is to remove from the anchor set the item with the largest d^2 and recompute the RSSS table. If there is a practical difference between the initial RSSS table and the second RSSS table at select points of interest (e.g., cut scores), then the item with the next largest d^2 value is dropped. This step-wise removal procedure continues until there are no practical changes in the RSSS tables across iterations. Such a procedure might be considered conservative in that it tends to lead to fewer items dropped.

The key objective of this study is to demonstrate a methodological procedure or strategy for examining the different anchor stability procedures and the accompanying results and to evaluate the impact on the final RSSS tables and reported cut scores (i.e., performance levels). For our study we did not include the bivariate plots for the old and new parameter values.

Method

Generated Data

The data for this research was generated using a real-data simulation technique. Item parameter estimates were collected from several of our testing programs that use the 3PL model and CINEG equating design in order to determine a reasonable set of values to use for modeling practical significance when simulating item parameter change. From these projects we chose the statistical characteristics of two programs to use for simulating realistic data sets. The scale transformation constants for the real programs were cataloged for generating Year 2 parameters (see Table 1). The absolute differences between Year 2 (postequated) and Year 1 real item parameter estimates for each program were computed. The maximum and standard deviation of these values were cataloged for generating realistic item parameter change and outliers among the Year 2 items (Table 2).

Table 1. IRT Scale Transformation Constants from Real Data

Test Number	A	B
1	0.96	0.03
2	1.07	0.03

Table 2. Maximum (Standard Deviation) of Absolute Item Parameter Differences from Real Data

Test Number	3PL Item Parameter		
	A	b	c
1	0.36 (0.08)	0.54 (0.09)	0.18 (0.03)
2	0.19 (0.05)	0.68 (0.18)	0.15 (0.03)

Year 1 item parameters and raw score distributions were generated using the program WinGen (Han, 2007) for test lengths of 30 and 50 items. All items appearing in Year 2 were considered anchors. Year 2 preequated item parameters were generated by adjusting the Year 1 values using the real data transformation constants from Table 1, plus a random noise factor intended to make the results more realistic:

$$\begin{aligned}
 a_{Year2} &= \frac{a_{Year1}}{A} + D_a R_a, & b_{Year2} &= Ab_{Year1} + B + D_b R_b, \\
 c_{Year2} &= c_{Year1} + D_c R_c
 \end{aligned} \tag{6}$$

where A and B are the real-data IRT scale transformation constants, D_x is the standard deviation of the real-data changes for each item parameter, and R_x is a random normal number drawn separately for each item parameter. If Formula 6 resulted in a negative value for the c -parameter, the Year 1 value was used instead.

To generate items with outlier item parameter changes we randomly selected one item from each Year 2 test to receive a “large” change. Large was defined as the outcome from Formula 6 plus or minus the maximum real item parameter change from Table 2. The sign was in the same direction as the adjustment made in Formula 6 in order to avoid having the effects cancel each other out. It was possible, though not probable, that an outlier item could be generated simply by Formula 6 without an adjustment. We introduced three conditions of change in item parameters: a) only b -parameter change, b) only a -parameter change, and c) both a - and b -parameter change. Since practitioners often ignore c -parameter changes we did not include this as a factor in this research. We acknowledge that in reality each of these parameters varies with one another, and therefore, we are departing somewhat from our goal of creating realistic changes. For each test we randomly selected one item to serve as the item with large change. The Year 1 and Year 2 base item parameters for this item is provided in for the selected items. The item parameters for Year 2 are previous to the adjustment for the large change which can be found in Table 2.

Table 3. Year 1 and Year 2 IRT parameters for Items Selected for Large Change – Before Maximum Change(s) From Table 2 Applied

Test Number	Number of Items		3PL Item Parameter		
			a	b	c
1	30	Year 1	0.696	1.308	0.284
		Year 2	0.679	1.202	0.327
	50	Year 1	1.238	1.348	0.227
		Year 2	1.331	1.388	0.236
2	30	Year 1	0.982	-0.226	0.420
		Year 2	1.020	-0.131	0.342
	50	Year 1	1.567	-0.409	0.148
		Year 2	1.569	-0.310	0.071

Note: The direction of maximum change(s) in a and/or b parameters in Year 2 can be determined by the difference between the Year 1 and Year 2 value. If the Year 2 value is greater than Year 1, the maximum change was added. If the Year 2 value was less than Year 1, the maximum change was subtracted.

We generated a Year 1 raw score to theta table for each test using the program POLYEQUATE (Kolen & Cui, 2004) and the Year 1 item parameters. We scaled the outcome theta values using a linear transformation with a slope of 35 and intercept of 600. This is a commonly used scale at Pearson, and reflects typical practice of transforming thetas to a reporting metric. We assigned raw score cuts at 40% and 75% correct, which are typical of state testing programs, and tabled the scaled scores associated with these raw score cuts. Theoretically, the noise introduced in Formula 6 should cancel out and after equating, and the Year 2 post-equated RSSS table should differ from the Year 1 table only to the extent that the item with large change contributes to the final results.

Equating Procedure

All items were considered anchors. The Stocking and Lord (1983) methodology was used to calculate the scale transformation constants to place the Year 2 item parameters on the Year 1 scale. The method has been used widely in large scale assessments (Baker & Al-Karni, 1991; Hanson & Beguin, 1999; Way & Tang, 1991). Stocking and Lord was implemented using STUIRT (Kim & Kolen, 2004). IRT true score equating was implemented using POLYEQUATE. The equating procedures were implemented identically for each anchor stability check approach:

1. Using the common items, obtain scaling constants using STUIRT.
2. Compute true score equating using POLYEQUATE (STUIRT provides transformed item parameter estimates that serve as inputs to POLYEQUATE).
3. Conduct the anchor stability check procedure (below) until the stopping rule is met. Note that we implemented the robust z procedure in a list-wise fashion, while the other procedures were conducted step-wise.
4. If Step 3 indicates an outlier and the remaining anchor set is greater than 80% of the total test length, drop the outlier from the anchor set and return to Step 1. Otherwise obtain final results.

The Year 2 raw score cuts were assigned by finding the post-equated raw score with the same scale score as the Year 1 cut. If an exact match was not found, then the raw score with the scale score closest to, yet greater than, the scale score cut was assigned.

Anchor Stability Check Methods Used

Although evaluation of new-old item parameter plots is a common methodology we chose not to use this method because the identification of an outlier is an arbitrary judgment. Instead we decided to identify possible inconsistencies between this approach and the others in our final analyses, if they exist.

IRT Parameter Differences. Differences in observed parameters were computed by subtracting the new parameter values from the old parameter values. Only the a - and b -parameters were considered. Items with absolute changes in magnitude larger than 0.30 for a or 0.50 for b were considered outliers. Although these criteria are arbitrary, we have used them in practice. We note that the 0.50 criteria for b is also a value sometimes used for Rasch equating. Stopping rule: if no items exceed 0.30 for a or 0.50 for b .

Robust Z. Robust z values were calculated via the following procedures, computed separately for the a - and b -parameters (*SCDE*, 2001):

1. Obtain the difference between new and old item parameters.
2. Calculate the mean, median, and interquartile range of the differences calculated in Step 1.
3. Calculate robust z .
4. Items with an absolute value of a robust z exceeding 1.645 for either parameter were considered outliers.
5. Stopping rule: if the robust z for no items exceed 1.645 for either the a - or b -parameter, or fewer than 80% of the test remains in the anchor set. Since this was accomplished in list-wise fashion it was possible that more items would be flagged than are allowed to be dropped. Items were rank ordered by magnitude of robust z , and those with the largest values were dropped.

*d*² Procedure. The steps for using the *d*² procedure were:

1. Apply the scale transformation constants to the thetas associated with the standard normal theta distribution.
2. For each anchor item calculate a weighted sum of the squared deviation between the ICCs based on old and new parameters at each point in a theoretically weighted posterior theta distribution
3. Sort the items in descending fashion according to *d*².
4. Flag the item with the largest *d*² as an outlier.
5. Stopping rule: if the raw score cuts do not change from the previous iteration. The scaling constants are the ones obtained from the previous iteration, which is the last iteration to have practical change.

Results

The numbers of items flagged by each anchor stability check method are provided in Table 4 and Table 5 for the 30 and 50 item test respectively. The robust z approach consistently flagged the item given the “large” change, but also flagged many more items, often exceeding our maximum drop rate of 20%. The *d*² approach flagged the next highest number of items. In the 30-item test it correctly flagged the item with “large” change for both tests and all conditions. However, in the 50-item test it only flagged the correct item in the both parameters changed condition of Test 1. The items flagged in the a-Change condition for Test 1 did not include the changed item. For the 50-item test, the *d*² approach flagged no items more often than any item. The item parameter difference approach only flagged the changed item in 5 of the 6 cases for the 30 item test (no item

was flagged in the a-Change condition for Test 2). For the 50-item test the changed item was flagged in the same conditions as with the 30-item test.

These results demonstrate the relative sensitivity of each procedure to changes in item parameter estimates. Given that the magnitude of changes we applied to the item parameters are commonly observed differences, the robust z approach (as we applied it) demonstrated over-flagging, while the d^2 and item parameter change approaches were more conservative in flagging, but sometimes did not flag the item with true change.

Table 4. Number of Items Flagged by Method and Condition - 30 Item Test

Test	a-Change			b-Change			Both		
	Δ	Robust z	d^2	Δ	Robust z	d^2	Δ	Robust z	d^2
1	1	9	1	1	7	3	1	8	1
2	0	6	5	1	6	1	1	6	1

Note: not more than 6 items could be dropped from the anchor set.

Table 5. Number of Items Flagged by Method and Condition - 50 Item Test

Test	a-Change			b-Change			Both		
	Δ	Robust z	d^2	Δ	Robust z	d^2	Δ	Robust z	d^2
1	1	14	2	1	12	0	1	11	1
2	1	10	0	2	9	0	2	9	0

Note: not more than 10 items could be dropped from the anchor set.

The differences between the true and estimated scale transformation constants from each method are provided in Table 6 and Table 7. If a particular approach flags only the item that was changed, we expect the estimated scaling constants to be recovered by the Stocking and Lord procedure. Recovery for the A-constant is the inverse of the true A value, while recovery for the B-constant is -1 times the true B value. Although the robust

z approach removed more items, and items that were not given a “true” change, it recovered the A constant well in these replications. In fact, the robust z approach recovered both A and B constants in the 30-item condition for Test 2 very well. B constant recovery was erratic for all three approaches. For the item parameter change and d^2 approaches, the final anchor sets differed only by no more than two items.

Table 6. True and Final Estimated Scale Transformation Constants by Method and Condition – 30 Item Test

Condition	Method	A Constant		B Constant	
		Test 1	Test 2	Test 1	Test 2
True	N/A	0.957	1.066	0.027	0.022
Recover Value	N/A	1.045	0.938	-0.027	-0.022
a -Change	Δ	1.044	0.925	-0.024	-0.011
	Robust z	1.044	0.942	-0.015	-0.022
	d^2	1.044	0.922	-0.024	-0.028
b -Change	Δ	1.044	0.913	-0.024	0.0001
	Robust z	1.044	0.942	-0.015	-0.022
	d^2	1.034	0.930	-0.024	-0.046
Both	Δ	1.044	0.913	-0.024	0.0001
	Robust z	1.044	0.942	-0.015	-0.022
	d^2	1.044	0.933	-0.024	-0.047

Table 7. True and Final Estimated Scale Transformation Constants by Method and Condition - 50 Item Test

Condition	Method	A Constant		B Constant	
		Test 1	Test 2	Test 1	Test 2
True	N/A	0.957	1.066	0.027	0.022
Recover Value	N/A	1.045	0.938	-0.027	-0.022

<i>a</i> -Change	Δ	1.038	0.937	-0.029	-0.021
	Robust z	1.044	0.936	-0.015	-0.047
	d^2	1.037	0.924	-0.033	-0.024
<i>b</i> -Change	Δ	1.038	0.930	-0.029	-0.014
	Robust z	1.044	0.943	-0.027	-0.043
	d^2	1.026	0.928	-0.037	-0.009
Both	Δ	1.038	0.930	-0.029	-0.014
	Robust z	1.035	0.943	-0.036	-0.043
	d^2	1.038	0.923	-0.029	-0.017

The root mean square difference (Harris & Crouse, 1993; Kolen & Harris, 1990) was computed:

$$\text{RMSD} = \left(\frac{\sum_i f_i (A_i - B_i)^2}{\sum_i f_i} \right)^{\frac{1}{2}} \quad (8)$$

where A_i is the raw score equivalent of the post-equated Year 2 scale i arrived at from the particular anchor stability check procedure, and B_i is the raw score equivalent of the Year 1 scale i , f_i is the frequency of a raw score of i on the Year 1 test; and i runs over the possible raw score range. The RMSD for each procedure and condition is provided in Table 8 and Table 9. **Error! Reference source not found.** We do expect differences between Year 1 and Year 2 tables due to the item that was given large change, and expected to be dropped from equating. Even so, we expected the procedures to flag that item, and to produce the same, or at least similar RMSD values. For both the 30- and 50-item conditions the differences between approaches for Test 2 were more pronounced than for Test 1.

Table 8. Root Mean Square Difference of Equating by Method and Condition - 30 Item Test

Condition	Method	Test 1	Test 2
<i>a</i> -Change	Δ	1.090	0.787
	Robust <i>z</i>	0.973	0.727
	d^2	1.090	1.035
<i>b</i> -Change	Δ	0.978	1.059
	Robust <i>z</i>	0.915	1.315
	d^2	1.140	0.817
Both	Δ	1.159	1.073
	Robust <i>z</i>	1.068	1.353
	d^2	1.159	0.868

Table 9 Root Mean Square Difference of Equating by Method and Condition - 50 Item Test

Condition	Method	Test 1	Test 2
<i>a</i> -Change	Δ	0.810	0.905
	Robust <i>z</i>	0.841	1.291
	d^2	0.810	1.026
<i>b</i> -Change	Δ	0.873	1.353
	Robust <i>z</i>	0.922	1.014
	d^2	0.817	0.879
Both	Δ	0.914	1.504
	Robust <i>z</i>	0.805	1.057
	d^2	0.914	1.014

Summary of changes in the final raw score cuts between the Year 1 and post-equated Year 2 tests for each procedure are found in Table 10. This shows changes, if

any, to cut scores. We note that for all cases, if a cut score changed after post-equating, it went up. This finding, though not expected, might be the result of a number factors, including that the Stocking and Lord recovery values were not completely retrieved, equating error, or slight bias in the random noise component.

Table 10. Raw Score Cut Changes for Equated Year 2 by Method and Condition

Condition	Method	30 Item Test				50 Item Test			
		Test 1		Test 2		Test 1		Test 2	
		C1	C2	C1	C2	C1	C2	C1	C2
<i>a</i> -Change	Δ	+1	+1	—	+1	—	—	+1	+1
	Robust <i>z</i>	+1	+1	+1	—	—	—	—	+1
	d^2	+1	+1	+1	+1	—	—	—	+1
<i>b</i> -Change	Δ	+1	+1	—	—	—	—	+1	+1
	Robust <i>z</i>	+1	+1	+1	—	—	—	—	+1
	d^2	+1	+1	+1	+1	—	—	—	+1
Both	Δ	+1	+1	—	—	—	—	+1	+1
	Robust <i>z</i>	+1	+1	+1	—	—	—	—	+1
	d^2	+1	+1	+1	+1	—	—	—	+1

Note: C1 stands for cut 1, and C2 for cut 2, — indicates no change

Discussion

Equating is integral part of maintaining the scale for most testing programs.

Ensuring that year-to-year item parameter stability is evaluated appropriately is core to the success of equating. While there is extensive research investigating different equating methods, research is sparse on the different methods available for conducting the anchor stability check. Since evaluating anchor stability is performed during equating in most large scale assessments, psychometricians have a need to know the how the different approaches for checking anchor stability might impact results. This study provides a strategy for investigating the various approaches.

We found that the approaches studied here provide very different results both in terms of the numbers of items flagged, and as a direct result, the final scale transformation constant estimates. The Robust z procedure flagged a relatively large number of items, typically resulting in implementation of a stopping rule other than the magnitude of the statistic itself. This fact rendered the magnitude of the statistic irrelevant as it was the secondary criterion that resulted in decision making. Simply rank ordering and dropping the top 20% of the items would have led to the same result. It is possible that it was our implementation of the procedure that led to this outcome. We chose a liberal flagging criterion of 1.645, and we dropped items if flagged for either the a - or b -parameter. The outcome here suggests additional research should be conducted to determine if a more conservative set of criteria would be more useful in practice.

We expected the item parameter change approach to be more conservative since it requires a constant threshold be exceeded for each parameter. Choice of the criterion values seems more arbitrary than the rules for the other procedures. The criterion we used *a priori* may be considered relatively large. In fact, these exceeded the maximum values that we observed in the real-world tests we used to model this study. Therefore, to be flagged, the item parameter change had to be very large even from our practical experience standpoint. From our perspective, this is concerning because we have observed quite a bit of variability in item parameter changes across projects. Using criterion that seems to work effectively on one project may be ineffective for another project. The approach may be conservative in one situation, and liberal in another. While the criterion could be tailored to the real-world observations of a particular program, this would require the kind of judgment on behalf of the equating analyst that we hope to

avoid in practice. In our opinion, this approach is too erratic to be effective for practical use.

The d^2 approach resulted in no-flagging in over $\frac{1}{2}$ of the 50-item test conditions, which suggests the procedure, as we implemented it, may be too conservative. This may be due to our additional rule that an item is dropped from the anchor set if an only if doing such leads to impact in the results. As of this time we do not have a criterion value for dropping items based on the statistic itself. We note that our use of the approach is straight forward for IRT true score equating, but is certainly not for projects that use ability estimation for score reporting. In the latter case, almost any change to the anchor set can be expected to have impact to some degree, and therefore, judgment is needed. The d^2 approach is theoretically appealing in that it focuses on changes to the item characteristic curves, not just one parameter at a time. It also weights the comparison based on the proportion of students theoretically scoring at each ability level. Changes that have little impact on the Stocking and Lord computation are not likely to result in rejection of an item.

Because we only evaluated the differences among procedures in two synthesized tests, this study is best considered a demonstration, and it is difficult to generalize beyond these findings. In our practice we routinely are asked to use different procedures and different criteria in implementing those procedures. Ideally, different procedures are used to triangulate on solutions. With respect to the ultimate outcome, equated cut scores, the results here were mixed. What is clear is that different outcomes can be observed when evaluating the same test with the different procedures. This demonstration also suggests that methods may not behave in a consistent manner, which would likely lead to more

systematic equating error. We strongly recommend that further research be conducted on these, and other, procedures. Additional research is also required that expands the number of synthesized tests and adds replications within these to understand how the anchor stability procedures operate in different testing situations.

Practitioners are often asked to produce equating results in a very short period of time, or to equate tests that have very high stakes, such as gate-keeper tests for high school graduation or grade promotion. In such cases practitioners are aided when defensible evaluative criteria are established so that the numbers of arbitrary judgments are reduced. Such criteria can lead to more automatable procedures and a reduction in opportunities for error. The item stability check procedures investigated here can be automated. Effective automation requires understanding of the sensitivity of the procedures to identify changes in item parameters under real-world conditions. We encourage researchers to pursue such solutions.

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.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.
- Hanson, B. A., & Beguin, A. A. (1999). Separate versus concurrent estimation of IRT item parameters in the common item equating design. *ACT Research Report Series, 99-8*, 1-34.
- Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195-240.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.
- Kim, S. & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional Item Response Theory models. The University of Iowa: Iowa City.
- 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).
- Kolen, M. J. & Cui, Zhongmin (2004). POLYEQUATE. The University of Iowa: Iowa City.
- Kolen, M. J., & Harris, D. J. (1990). Comparison of item pre-equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*, 27-39.

- Miller, G. E., & Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educational and Psychological Measurement, 69*, 357-368
- South Carolina Department of Education. (2001). *Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics* (Technical Report). Columbia: South Carolina Department of Education.
- 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.