MCAS Equating Research Report: An Investigation of FCIP-1, FCIP-2, and Stocking and Lord Equating Methods[1,2]

Lisa A. Keller, Ronald K. Hambleton, Pauline Parker, Jenna Copella

University of Massachusetts Amherst

December 10, 2008

---

[1] Center for Educational Assessment MCAS Validity Report No. 19. (CEA-690). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

[2] This research was carried out for the Massachusetts Department of Elementary and Secondary Education (MDESE) under a contract between the Center for Educational Assessment at the University of Massachusetts Amherst (Ronald Hambleton, PI) and Measured Progress (MP), Dover, NH. We are grateful to the MDESE, their technical advisory committee, and MP for their many suggestions for conducting the study and for their assistance in interpreting the findings.

Table of Contents

Executive Summary 3

I. Introduction to the Report..... 6

II. Proposed Research Plan 6

III. Study 1: A Simulation Study to Compare FCIP-1, FCIP-2, and Stocking & Lord Equating Methods Over a Four-Year Period 12

IV. Study 2: A Simulation Study to Evaluate the Effect of Changing Equating Methods 22

V. Study 3: A Simulation Study to Investigate the Effect of Test Difficulty on Equating Accuracy.....25

VI. Summary of Results from the Three Studies.....40

References.....42

Executive Summary

Due to recent research in equating methodologies that have indicated that some methods may be less accurate in capturing growth in examinee ability distributions, the Massachusetts Department of Education commissioned research into the accuracy and sustainability of various equating methods in the context of the Massachusetts Comprehensive Assessment System (MCAS). Three studies were conducted to compare three equating methods: two implementations of fixed common item parameter equating, FCIP-1 and FCIP-2, and the Stocking and Lord test characteristic curve method. A brief summary of the three studies is provided next.

In the first study we investigated the sustainability of the three methods across four administrations of the test. The effect of changing methods from the FCIP-1 to FCIP-2 method was investigated in the second study. In the third study the effects of the difficulty of the test on the quality of the equating were investigated. For each of the three studies, the same data simulation plan was followed, and the same evaluation criteria were used.

Data simulation included simulating response data to items that had the same item parameters as the operational MCAS items. The test format (29 binary-scored multiple choice items, 4 binary-scored short answer items, and 5 constructed response items (scored 0 to 4) was chosen to mirror that of the grade 6 MCAS mathematics test, which was determined to be representative of the various types of tests in the MCAS. Examinee ability distributions were manipulated to represent two different growth patterns: Consistent growth where all examinees increased in ability at the same rate between administrations, and differential growth where examinees at the lower end of the distribution exhibited more growth than examinees in the upper end of the distribution. In both cases, the empirical ability distribution from a grade 6 mathematics test was chosen for the base ability distribution and was manipulated.

The equating methods were evaluated based on two main criteria: The amount of growth captured between administrations and the classification of examinees into performance categories. The amount of growth captured was calculated as a change in the mean ability between two subsequent administrations, and this value was compared to the amount of growth that was simulated. The classification of examinees was examined by looking at the number of examinees that were correctly classified into the four performance categories, the number of examinees that were classified into a lower category than they should have been (underclassified) and the number of examinees that were classified into a higher category than they should have been (overclassified). The results were examined at the overall level, for all examinees, and also at the conditional level, where the classification accuracy was evaluated for examinees at each true performance level. For example, the classification of examinees whose true classification was Proficient was examined separately from the other examinees. This analysis was repeated for each performance category.

Results from Study 1 indicated that the FCIP-2 and the Stocking and Lord methods produced more accurate results than the FCIP-1 method (1) in assessing growth (both consistent and differential) and (2) in assigning examinees correctly to performance categories. Regarding the latter, the difference ranged from 2 to 4% depending on the year, which translated to a difference in accuracy of classification for somewhere between 1,400 and 2,800 examinees (with a total of 70,000 examinees in the sample). When the Proficient category is considered, as this is arguably the most important for No Child Left Behind purposes, there is a difference of about 3% of correct classifications between the FCIP-1 method and either the Stocking and Lord or the FCIP-2 method. A 3% difference in classification accuracy translates into more accurate classification for about 2,100 examinees. This is a practically significant difference and could easily be used to justify either of the better equating methods.

Study 2 was designed to investigate the effect of changing methods from FCIP-1 to FCIP-2. The results of the study indicated that when the change was made, generally, there was less error in the estimate of growth and the classification of examinees. The improvement was even more substantial for the consistent growth condition than for the differential growth condition. The difference in classification accuracy was about 3%, which again translated to about 2,100 examinees in the state. As such, the change would result in an increase of about 2,100 examinees being correctly classified.

The results of Study 3 indicated that in the case where the anchor items were easier than the scoring items (Case A), the FCIP-2 method produced the best estimate of the true growth. For overall classification rates, the Stocking and Lord and FCIP-2 methods produced similar results, and both produced greater accuracy of examinee classification than the FCIP-1 method. These are important findings because Case A is a common scenario in practice—it is the case where as examinees demonstrate more proficiency over time, test difficulty is revised accordingly to improve the precision with which examinee proficiency is estimated. In this case too, the difficulty of the equating items remains about the same from year to year. Findings from Case B were similar to those obtained with Case A, but this case is not realistic in practice—test developers would rarely if ever construct relatively easier tests each year when the examinee proficiency distribution is increasing.

The overall trends in the findings in these three studies indicate that the FCIP-2 and Stocking and Lord methods performed similarly to each other, and produced more accurate estimates of growth, and more accurate classification of examinees than the FCIP-1 method. The differences do not seem great (typically between 1 and 4%), but they are of considerable practical consequence, and the FCIP-2 method could be implemented with only a minor change or two in the current equating process in Massachusetts. In terms of changing methods from FCIP-1 to FCIP-2, the results of Study 2 indicated that when the change was made, there was a reduction in the error of the estimation of growth and an improvement in the classification accuracy of examinees. While the changes may not seem dramatic, due to the large sample sizes in the MCAS program, as noted previously, differences of even 3% can affect the performance classifications of a large number of examinees—about 2,100 examinees.

Therefore, taken together, the overall results of these three studies indicate that using either FCIP-2 or Stocking and Lord equating methods appear to lead to somewhat more accuracy than the FCIP-1 method. As the simulations (with the exception of Case B in Study 3) very much match the tests and the conditions in the MCAS program from year to year, generalization of the main findings to other grades 3 to 8 ELA and mathematics tests would seem to be warranted. The results from the research based on theoretical support (it is best to reset the prior proficiency distribution after each stage in item parameter estimation to reflect the best available estimates of the mean and standard deviation of the proficiency scores) as well as the empirical evidence from the three studies, strongly support a decision to switch from the FCIP-1 to the FCIP-2 equating method.

I. Introduction to the Report

This report is a summary of the results of the Massachusetts Comprehensive Assessment System (MCAS) equating studies in 2007-2008 conducted at the University of Massachusetts Amherst for the Massachusetts Department of Education. The report begins with an introduction and a revised version of the proposal that was presented to the Massachusetts Department of Education and the MCAS Technical Advisory Committee in January of 2008. The revisions to the proposal reflect the decisions that were made at the MCAS Technical Advisory Committee Meeting. The third section provides the results of the first study. The fourth section presents the results of the second study, and the fifth section presents the results of the third study. After the presentation of the three studies separately, a summary of the results of all the studies is presented, followed by a discussion of the results.

II. Proposed Research Plan

Questions are always being asked by psychometricians about whether better equating methods can be found to capture or estimate the changes in student performance over time (see, for example, Jodoin, Keller, & Swaminathan, 2003), and as such, it is essential to be certain that the equating methods and designs chosen to equate statewide test results are the very best choices possible for the test design and the purposes of testing. Since one of the goals of the MCAS is to measure progress of students as defined by the number of students classified into performance categories, it is important that the results of the equating are evaluated with regards to the accuracy of the classification of the students into performance categories. Several studies have already documented the effect of the equating method on the estimation of student and item parameters, however, to evaluate how practical the differences between methods are, it was important to consider the effects on the classification of students. In addition to the classification of students, it was also informative to investigate the accuracy with which changes in student performance are captured.

Three studies were proposed to investigate three equating methodologies: FCIP-1, FCIP-2, and Stocking and Lord. Additionally, investigating the effect of changing from FCIP-1 to FCIP-2 was proposed. The proposed research project consisted of three studies: Study 1: A Simulation Study to Compare FCIP-1, FCIP-2, and Stocking & Lord Equating Methods over a Four Year Period, Study 2: An Investigation of the Effect of Changing Equating Methods, and Study 3: The Effect of Test Difficulty on Equating Accuracy. There has been a paucity of research regarding the sustainability of equating methodologies over multiple administrations, and hence to be certain that the chosen method can withstand long-term use, research in this area was considered essential. The equating methodologies that were investigated were two implementations of FCIP (called here, FCIP-1 and FCIP-2) and the Stocking and Lord (S-L) Test Characteristic Curve (TCC) method. Brief descriptions of the FCIP methods are presented next.

Fixed Common Item Parameter Equating

In the fixed common item parameter (FCIP) equating methodology, the item parameters for the anchor items are fixed in the calibration process to the values obtained from a previous year's administration, and treated as known. The parameters of the remaining (non-anchor) items are estimated and placed onto the scale of the anchor item parameters, hence accomplishing the equating. Kim (2006) considered various ways to implement the FCIP method by making refinements to the estimation procedure. His study showed that there were differences in some of the implementations of FCIP when the difference in performance in the two groups being equated was large: At least one half of a standard deviation. Of course, the practical consequences of such differences should be ascertained, as little is known about the effect of the various methods on the classification of students in a realistic context.

Only two of the five FCIP methods investigated by Kim (2006) are relevant for discussion, as the others cannot be conveniently implemented with the current commercial software. The differences between the two methods are related to the way that the prior distribution of the proficiency parameter is handled in the estimation. Since PARSCALE, and most of the popular IRT packages, use marginal maximum likelihood estimation of the item parameters, it is necessary to specify a distribution for the performance parameter. The default is to place a N(0,1) distribution on the ability parameter. After each EM-cycle, the prior distribution can be updated, or not. It is theoretically more sensible to update the prior distribution each time, which is what FCIP-2 does, while FCIP-1 does not. Hence, it is not surprising that the two methods perform differently, at least in some contexts. It was necessary to consider the performance of the FCIP methods when used over multiple years, and under more realistic contexts. To differentiate the methods of implementing FCIP, FCIP-1 will refer to the typical way that FCIP is performed, and FCIP-2 will refer to the adjustment suggested by Kim, which, in his study, led to improvements in the estimation of proficiency estimation when growth was present from one test administration to the next.

Keller, Keller, and Baldwin (2007) did a preliminary investigation of the effectiveness of several equating methodologies, including FCIP-1 and FCIP-2, in a multiple-administration context. They found that while the results of FCIP-2 did seem to be an improvement over those of FCIP-1, the results seemed to show some breakdown in the process when used over a longer period of time. Further, the shape of the performance distribution seems to have an impact, with skewed distributions being slightly more problematic than symmetric (or at least normal) distributions. Therefore, based at least on this preliminary study, and these limited results, the FCIP-1 and FCIP-2 equating methodologies bear additional investigation. Keller and Keller (2007a) are doing a more detailed study to investigate the effectiveness of the FCIP methods over multiple administrations as well as investigating the effect of changing methods (Keller & Keller, 2007b). These studies utilize simulation studies that were not directly analogous to the MCAS testing situation. Specifically, a matrix-sampled design was not used, which may have at least some effect on any findings. Therefore, the goal was to design this study to match the testing program in Massachusetts as closely as possible,

especially in terms of test design. Performance distributions were chosen from the score distributions observed in recent years.

Using simulated data requires that the simulation of the data model MCAS as closely as possible to generalize to operational MCAS results. To ensure that the generalization was appropriate, operational item and person parameters were used to generate the simulated student response data. Therefore, before describing the three studies in detail, general details that apply to all three studies are provided first.

Test Design

The design of the simulated tests mirrored the MCAS test design in terms of the number and types of items. The matrix-sampled structure was maintained as well as the proportion of dichotomously-scored and polytomously scored items. Given that the format of the test changes somewhat from subject to subject, and grade to grade, however, it was impossible to mirror the design of all possible MCAS tests exactly. As such, it was decided to choose a test design that mirrors a typical test using FCIP equating, say, 6^{th} grade mathematics, since it was in the middle of the grades where FCIP equating is used (grades 3 to 8). In the absence of any compelling evidence for one choice over another, and given the need to constrain this study to a practical size, we limited our attention to the results from one test at one grade over several administrations. It should be noted that the choice of grade and subject matter was arbitrary, and that the choice would affect the number of items included on a form, and the particular item parameters. But within that framework, many important variables were manipulated, and their effects studied.

Generating Data

When generating data, there are three primary aspects to consider:

(1) Generating the performance (i.e., proficiency) parameters

(2) Generating the item parameters

(3) Generating responses to items

Each of these will be described next.

Generating Performance (Proficiency) Parameters

Performance parameters were generated according to the empirical distribution of the operational performance parameters. Again, as there are many grades and subjects, there were many potential distributions that could be modeled. Similar to the test design issue, we chose a middle level grade—grade 6. Again, we chose mathematics. Performance parameters were simulated for the number of students that are typically administered the test in a given year (about 70,000). Given that the studies proposed

span several years, changes in the performance distribution were expected, and those changes were modeled. Two types of changes were modeled: Consistent growth, and differential growth.

Consistent Growth: Using the empirical distribution as the year 1 distribution, a shift in the mean (keeping the shape of the distribution constant) was modeled by increasing the mean by .15 standard deviations each year for three additional years. A shift of this magnitude corresponds to a difference of 2 to 3 test score points (on average) each year.

Differential Growth: Again, using the empirical distribution as the year 1 distribution, the next three distributions were obtained by shifting the skewness of the distribution moderately each year. This was accomplished by changing the amount of growth for certain quantiles of the distribution. Therefore, examinees in the lower end of the distribution changed more than examinees in the upper end of the distribution. Specifically, the following steps were taken to simulate the desired growth:

1. The examinees in year 1 were broken into five groups based on their theta (proficiency) parameters. Group 1 represented the group of lowest ability and group 5 represented the group of highest ability.

2. Examinees in group 1 were shifted by .30, in group 2, by .20, in group 3 by .15, in group 4 by .10 and in group 5 by .05. This resulted in the type of shift that was sought to investigate. (We are not sure how realistic this case is, but wanted to simulate a case where the lower performing students were getting lots of instructional assistance and were showing more growth from year to year than the more able students.)

This same process was repeated for each administration.

Zero to .45 SD changes in the means were used in our study because the results of Kim (2006) were limited to the case where the differences in means were very large (about .5 of a standard deviation). It is important to understand the results of equating when the gains are much smaller than .5 of a SD, as they are likely to be less than .5 SDs in many operational testing programs.

Generating Item Parameters

Operational item parameters were used in the study to simulate the item response data. Parameters from grade 6 mathematics were used. Since the studies involved multiple administrations, the item parameters for the first administration were based on the operational item parameters, and for each subsequent administration they were manipulated. Three different scenarios were investigated: (1) the item parameters for both the scoring and anchor items remained the same difficulty from year to year (the most common situation in practice), (2) the scoring items became more difficult each year, to match the changes in the ability of the examinees, while the difficulty of the anchor items remained of equal difficulty across administrations, or (3) the scoring items were easier

than the anchor items, which remained of approximately equal difficulty across administrations (this situation is not realistic in practice, and in hindsight it was probably not a good choice to make).

Generating Item Responses

Item responses were based on the IRT model used operationally to model the items. Therefore, the three-parameter logistic (3PL) model was used to simulate responses to multiple-choice items, the two-parameter logistic (2PL) model was used to simulate responses to the short answer (dichotomously scored) items, and the graded response model (GRM) was used to simulate responses to the five point (0 to 4) constructed response items. Using the appropriate model, the item parameters and the person parameters, the responses to all items were generated according to the model-based probabilities.

Replications

Given the large samples of examinees that respond to the test, the role of sampling variability was expected to be minimal. However, to reduce the effect of sampling variability on the results of the studies, ten replications were conducted for each simulation study and the averages of the results were reported. All results reported later in the studies are the averages over 10 replications.

These general guidelines were used in each of the proposed simulation studies. The details of each of the studies are presented after a brief discussion of criteria for evaluating the findings.

Evaluation Criteria

Five criteria were used to evaluate the equating methods under various conditions. These are (1) the difference in the observed growth as compared to the simulated growth, (2) the percent of candidates sorted into the correct performance categories, (3) the percent of students misclassified into a category higher than they should have been (over-classified), (4) the percent of students classified into a category lower than they should have been classified (under-classified) (5) the classification results conditional on the true performance of the examinees. Details of each of these measures are described next.

Criterion one: The differences between the observed growth and the simulated growth. The error in the growth estimates was calculated for each method for each of the years 2, 3 and 4. This was accomplished using the following formula:

$$\text{Error} = \Delta_{\hat{\theta}} - \Delta_{\theta}$$

where Δ_{θ} is the true growth between administrations, $\Delta_{\hat{\theta}}$ is the average estimated growth. Given this formulation of error, a negative value would represent an underestimate of growth, and a positive value would represent an overestimation of growth.

Criteria 2 through 5: Classification accuracy. After the various equatings were accomplished, the simulated examinees were placed into one of four performance categories based on the 2006 math grade 6 operational cut scores. This process was repeated for each of the equating methods across each of the years 1, 2, 3 and 4. The number of examinees that were classified into each of the performance categories based on both their true performance and their estimated performance was then tabulated. For example, the number of students that were classified as proficient based on their true performance and classified as Advanced based on their estimated performance was tabulated. This was done for all performance levels. Therefore, there were 16 possible categories into which an examinee could be classified, based on the 4 x 4 contingency table shown below:

An Example of a 4 x 4 Contingency Table For Classifying Examinees

		True Classification			
		Warning	Needs Improvement	Proficient	Advanced
Estimated Classification	Warning	Correct	Under	Under	Under
	Needs Improvement	Over	Correct	Under	Under
	Proficient	Over	Over	Correct	Under
	Advanced	Over	Over	Over	Correct

Since each simulated examinee was classified based on their true performance level and their estimated performance level, each simulated examinee would be classified into one of the sixteen cells in the contingency table. The percent of students in each of the sixteen cells was computed. The overall accuracy was computed by taking the sum of the diagonal elements of the table. Under-classifications are represented by the upper triangle of the table, while over-classifications are represented by the lower triangle of the table. Therefore, for each year and each method, the results were presented for the correct classification, the percent of over-classifications and the percent of under-classifications. The results were then broken down to look at the classification results conditional on the true performance level of examinees. Therefore, the same results were presented separately for the examinees in each of the performance levels, to examine if the misclassifications occur at some levels at a greater rate than other levels. Again, this was repeated for each method, and each year.

III. Study 1: A Simulation Study to Compare FCIP-1, FCIP-2, and Stocking & Lord Equating Methods Over a Four-Year Period

Four administrations of the MCAS were simulated using the grade 6 MCAS mathematics item parameters and performance parameters. In year 1, the item parameters from the 2006 administration of MCAS were used along with the performance parameters from that administration to simulate item responses to the items. The test design, as described above, was kept the same as the operational test design, with the matrix-sampling of the equating items. Five forms were simulated and anchor items were randomly assigned to one of the five forms. Each form was composed of 39 scoring items: 29 were multiple-choice items, 5 were short answer items, and 5 were polytomous items for a total of 54 operational scoring points. Additionally each form had 8 equating items: 6 multiple-choice items, 1 short answer item, and 1 constructed response item. The short answer items were scored dichotomously, and the constructed response items were scored on a 5 point scale (0 to 4).

Subsequent simulated administrations were simulated by constructing forms from field test items from 2006 and all items from 2007, to make forms that were as parallel as possible. The item composition of the subsequent administrations was identical to that of the first administration. Descriptive statistics for the item parameters can be found in Table 1 below.

Table 1
Mean Parameters for Each Administration by Item Type

Type	Parameter	Year 1	Year 2	Year 3	Year 4
Scoring: Dichotomous	a	0.93	0.98	0.93	0.99
	b	-0.34	-0.34	-0.35	-0.36
	c	0.16	0.16	0.14	0.16
Anchor: Dichotomous	a	0.97	0.97	0.97	0.97
	b	-0.35	-0.35	-0.35	-0.35
	c	0.16	0.16	0.16	0.16
Scoring: Polytomous	a	1.00	1.11	1.19	1.18
	b	-0.06	-0.11	-0.11	-0.12
Anchor: Polytomous	a	0.93	0.93	0.93	0.93
	b	-0.06	-0.12	-0.12	-0.10

It should be noted that while the test forms remained of the same difficulty from year to year, as the examinees became more proficient, the tests appeared easier to the examinees. So, while the test were constructed to be as parallel as possible, they would appear easier to the examinees as they became more able.

Performance parameters were manipulated to simulate growth in the examinee population. The distribution of the parameters was changed across the four administrations in two ways: Either all examinees exhibited the same amount of growth

between administrations, or the examinees exhibited differential growth between administrations. In the case where all examinees exhibited the same amount of growth, the mean of the proficiency distribution was shifted by .15 in each administration. In the case where there was differential growth, the examinees at the bottom end of the distribution were simulated to show more growth than the examinees in the upper end of the distribution, to model what might be expected to happen in practice, with the No Child Left Behind law, and instructional focus on the lower performing examinees.

The descriptive statistics for the performance parameters for both cases are presented below. Year 1 statistics are from the empirical distribution of the examinees, and were not simulated (the year one mean was .30 with an SD of .89). Therefore, the mean, SD and skewness values represented the empirical distribution of examinees. Because the performance distribution of the examinees started as negatively skewed, in the case of differential growth, the skewness of the distribution changed somewhat, and actually became less skewed, as the bottom tail of the distribution shrank, making the distribution more symmetric.

Table 2
Descriptive Statistics of Performance (Proficiency) Parameters

		Mean	SD	Skewness
Consistent Growth Condition	Year 1	.30	.89	-.42
	Year 2	.45	.89	-.42
	Year 3	.60	.89	-.42
	Year 4	.75	.89	-.42
Differential Growth Condition	Year 1	.30	.89	-.42
	Year 2	.46	.85	-.30
	Year 3	.54	.81	-.25
	Year 4	.62	.78	-.20

Table 3 below provides the number of students with true classifications in each of the four performance categories for both the consistent and differential growth categories. Note that the Year 1 values are identical for the two conditions, as the growth was not introduced until Year 2.

Table 3
Percent of Students in Each Performance Category

Growth	Year	Performance Category			
		Warning	Needs Improvement	Proficient	Advanced
Consistent Growth	Year 1	17.7	25.9	38.5	17.8
	Year 2	14.3	22.8	39.6	23.3
	Year 3	11.4	19.8	39.2	29.6
	Year 4	8.9	17.1	37.6	36.4
Differential Growth	Year 1	17.7	25.9	38.5	17.8
	Year 2	14.3	22.8	39.6	23.3
	Year 3	10.8	22.8	40.7	25.7
	Year 4	7.7	20.1	43.8	28.4

Ten replications of each of the administrations were simulated and the results were averaged to reduce the error in the findings due to sampling variability alone. The four administrations were then equated with each of the three scaling methods proposed: FCIP-1, FCIP-2, and Stocking and Lord.

Results

Determination of the Amount of Growth Captured

The results for the error in the growth estimates for both the consistent and differential growth conditions are presented in Table 4.

Table 4
Error in Growth Estimates for All Equating Methods, for Each Year

	Simulated Growth			Examinee Growth Conditions	
Year	Consistent	Differential	Equating Method	Consistent	Diffferential
Year 2	.15	.15	SL	0.05	0.05
			FCIP-1	-0.08	-0.08
			FCIP-2	0.00	-0.01
Year 3	.15	.08	SL	0.01	-0.06
			FCIP-1	-0.04	-0.09
			FCIP-2	-0.02	-0.08
Year 4	.15	.08	SL	0.02	-0.05
			FCIP-1	-0.05	-0.17
			FCIP-2	0.00	-0.05

The results for the error in the growth estimates show that the FCIP methods led to generally underestimates of growth, and the Stocking and Lord method overestimating growth, in the consistent growth condition. In the case of differential growth, all methods tended to underestimate growth. In all of the analyses reported in Table 4, the FCIP-1 produced the poorest results (i.e., these results were the furthest from the true indication of growth).

Overall Classification Accuracy Results

The results for the overall decision accuracy are presented first in Tables 5 and 6. Table 5 represents the overall results for the case where the simulated growth was consistent. Table 6 corresponds to the case where there was differential growth. It should be noted in Year 1, the only equating is to place the parameters back onto the generating scale, and as such, there are not different methods to do so. These results are included in the table to provide a baseline for the accuracy of the classification of the examinees. However, all the classification accuracy results are dependent upon several factors that are independent of the examinees themselves, such as the placement of the cut scores relative to the distribution of the examinees and the shape of the test information function. Therefore, comparisons across years are difficult to interpret, as the shape of the examinee performance distribution changes across years. An increase/decrease in accuracy from one year to the next might merely be due to the relative number of examinees near a cut point. Therefore, when examining the results of the classification accuracy, comparing methods *within* a year are quite meaningful, but looking across years could be problematic due to confounding factors.

Table 5
Percent of Correctly Classified, Over-classified and Under-classified Examinees, Consistent Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.70	0.16	0.13
	FCIP-1	0.69	0.11	0.20
	FCIP-2	0.71	0.14	0.16
Year 3	SL	0.71	0.16	0.13
	FCIP-1	0.68	0.10	0.22
	FCIP-2	0.71	0.13	0.16
Year 4	SL	0.73	0.16	0.11
	FCIP-1	0.71	0.11	0.18
	FCIP-2	0.73	0.12	0.15

In the case where the anchor and scoring items were of similar difficulty, it is apparent that FCIP-2 and Stocking and Lord led to almost identical results, while FCIP-1 had fewer examinees classified correctly. The only difference between FCIP-2 and Stocking and Lord results occurred in the second year, where FCIP-2 had 1% more students accurately classified. FCIP-1 and FCIP-2 typically had misclassifications that were under-classifications, while the misclassifications resulting from Stocking and Lord were typically over-classifications.

Table 6
Percent of Correctly Classified, Over-classified and Under-classified Examinees, Differential Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.70	0.16	0.14
	FCIP-1	0.68	0.11	0.21
	FCIP-2	0.70	0.14	0.16
Year 3	SL	0.69	0.17	0.14
	FCIP-1	0.66	0.10	0.24
	FCIP-2	0.69	0.14	0.18
Year 4	SL	0.71	0.16	0.13
	FCIP-1	0.68	0.12	0.21
	FCIP-2	0.70	0.13	0.17

As in the case of the consistent growth, the pattern of results suggested that FCIP-2 and Stocking and Lord performed similarly, with the FCIP-1 method producing less accurate classifications. FCIP-2 and Stocking and Lord showed a 1% difference again, this time in Year 4, with the Stocking and Lord method producing slightly more correct classifications. As in the case of consistent growth, Stocking and Lord tended to over-classify examinees, while FCIP-1 and FCIP-2 tended to under-classify examinees.

Conditional Classification Accuracy Results

The classification accuracy results were then broken down into separate results for examinees in each performance category. Tables 7 to 10 present the results for the consistent growth condition. Tables 11 to 14 present the results for the differential growth condition. It should be noted in all cases, where there was a misclassification, the misclassification was into an adjacent performance level. The misclassification never exceeded more than one category. Thus, if the true performance level classification for an examinee was Proficient, a misclassification would be either Needs Improvement, or Advanced, but never Warning. Each table will be presented in turn, followed by a summary of the results in that table.

Table 7
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Consistent Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.81	0.21	X
	FCIP-1	0.87	0.15	X
	FCIP-2	0.84	0.18	X
Year 3	SL	0.81	0.22	X
	FCIP-1	0.88	0.16	X
	FCIP-2	0.84	0.19	X
Year 4	SL	0.76	0.22	X
	FCIP-1	0.82	0.17	X
	FCIP-2	0.80	0.20	X

When the true performance level of the examinee was Warning, there clearly can be no under-classification of the examinee. Therefore, all the misclassifications were over-classifications, where the examinee was wrongly classified as being in the Needs Improvement category. Comparing methods, it was apparent that the FCIP-1 method led to the greatest accuracy in classifying the Warning examinees, and Stocking and Lord led to the greatest errors in classification. FCIP-2 fell between these two methods. This result was expected because the FCIP-1 method was found in other analyses to underestimate growth, and so it was expected that the FCIP-1 method was more likely to result in "true" Warning students being classified as Warning students.

Table 8
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Consistent Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.60	0.25	0.14
	FCIP-1	0.62	0.17	0.20
	FCIP-2	0.61	0.22	0.16
Year 3	SL	0.60	0.26	0.13
	FCIP-1	0.62	0.17	0.20
	FCIP-2	0.62	0.22	0.16
Year 4	SL	0.61	0.28	0.11
	FCIP-1	0.64	0.20	0.17
	FCIP-2	0.64	0.22	0.14

When the true classification of the examinee was Needs Improvement, there was a similar trend, where FCIP-1 had the most accurate results and Stocking and Lord were the least accurate. However, in this case, FCIP-1 and FCIP-2 were almost identical, with only a difference in Year 2, where there was a 1% difference between the two. In this case, there was the potential for both over-classification (classified as Proficient) and under-classification (classified as Warning). Stocking and Lord and FCIP-2 tended to over-classify examinees when there was a misclassification, while FCIP-1 tended to under-classify examinees when there was a misclassification.

Table 9
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Consistent Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.66	0.19	0.14
	FCIP-1	0.66	0.12	0.21
	FCIP-2	0.67	0.16	0.16
Year 3	SL	0.66	0.22	0.13
	FCIP-1	0.66	0.12	0.22
	FCIP-2	0.68	0.16	0.16
Year 4	SL	0.64	0.24	0.11
	FCIP-1	0.66	0.16	0.17
	FCIP-2	0.67	0.17	0.14

For students whose true classification was Proficient, FCIP-2 led to the most accurate classification of examinees, with Stocking and Lord and FCIP-1 showing similar rates of correct classification. Again, Stocking and Lord tended to over-classify students in the event of a misclassification, while FCIP-1 tended to under-classify examinees. FCIP-2 produced balanced errors in Years 2 and 3, and a greater rate of over-classification in Year 4.

Table 10
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Consistent Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.82	X	0.19
	FCIP-1	0.71	X	0.30
	FCIP-2	0.78	X	0.23
Year 3	SL	0.82	X	0.17
	FCIP-1	0.67	X	0.31
	FCIP-2	0.76	X	0.22
Year 4	SL	0.87	X	0.14
	FCIP-1	0.78	X	0.23
	FCIP-2	0.81	X	0.19

For the case where the true classification of examinees was Advanced, there was a change in the pattern of results. In this case, the Stocking and Lord method led to dramatically more correct classifications than either FCIP-1 or FCIP-2. FCIP-1 led to the least accurate results, with a substantial number of misclassifications. In this instance, any misclassification would be an under-classification, as there was no category higher than Advanced.

In summary, the different methods performed differently depending on the true classification of the examinee. For the examinees at the lower end of the distribution, FCIP-1 led to the most accurate results, while for the upper end of the distribution, Stocking and Lord method led to the most accurate results. The methods were fairly similar in all cases except when the examinee true classification was Advanced, and then the FCIP-1 method had very low rates for correct classification.

The conditional results for the differential growth condition are presented next in Tables 11 to 14. Again each table is presented followed by a summary of the results for that table.

Table 11
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Differential Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.75	0.25	X
	FCIP-1	0.84	0.16	X
	FCIP-2	0.78	0.22	X
Year 3	SL	0.70	0.30	X
	FCIP-1	0.81	0.19	X
	FCIP-2	0.76	0.24	X
Year 4	SL	0.69	0.31	X
	FCIP-1	0.79	0.21	X
	FCIP-2	0.76	0.24	X
	FCIP-1 to 2	0.77	0.23	X

In the differential growth condition, as in the case of consistent growth, when the true classification of the examinee was Warning, FCIP-1 led to the most accurate results. Stocking and Lord method led to the least accurate results, and FCIP-2 results were between the two. For Year 4, FCIP-1 produced dramatically better (10%) results than Stocking and Lord. Of course, in this case, all misclassifications were over-classifications.

Table 12
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Differential Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.62	0.24	0.14
	FCIP-1	0.62	0.16	0.22
	FCIP-2	0.62	0.22	0.16
Year 3	SL	0.60	0.28	0.12
	FCIP-1	0.60	0.17	0.23
	FCIP-2	0.61	0.23	0.16
Year 4	SL	0.61	0.27	0.12
	FCIP-1	0.59	0.20	0.21
	FCIP-2	0.61	0.22	0.17

When the true classification of the examinees was Needs Improvement, in the differential growth condition, the differences between the methods in terms of the percent of correct classifications was small (1 to 2%). Generally speaking, when there was a misclassification, for Stocking and Lord and FCIP-2, the misclassification was slightly more likely to be an over-classification, while for the FCIP-1 method it was more likely to be an under-classification.

Table 13
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Differential Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.68	0.18	0.14
	FCIP-1	0.66	0.12	0.22
	FCIP-2	0.68	0.15	0.17
Year 3	SL	0.68	0.18	0.14
	FCIP-1	0.65	0.11	0.24
	FCIP-2	0.68	0.15	0.17
Year 4	SL	0.69	0.18	0.13
	FCIP-1	0.65	0.14	0.21
	FCIP-2	0.68	0.15	0.17

For examinees with a true classification of Proficient, again, the differences in correct classification between the methods ranged from 2% to 4%. Stocking and Lord and FCIP-2 methods produced almost identical results, and were better than those obtained with FCIP-1. As before, the percent of over-classifications was greater than under-classifications for Stocking and Lord while for FCIP-1 and FCIP-2, the misclassifications were slightly more likely to be under-classifications.

Table 14
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Differential Growth

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.78	X	0.22
	FCIP-1	0.69	X	0.31
	FCIP-2	0.75	X	0.25
Year 3	SL	0.78	X	0.22
	FCIP-1	0.65	X	0.35
	FCIP-2	0.74	X	0.26
Year 4	SL	0.81	X	0.19
	FCIP-1	0.74	X	0.26
	FCIP-2	0.78	X	0.22

For students with a true classification of Advanced, the same pattern was observed as for the consistent growth condition, with Stocking and Lord producing the most accurate classifications, and FCIP-1 the least accurate, with FCIP-2 in the middle. The differences between FCIP-2 and Stocking and Lord were around 3% while the differences between FCIP-1 and Stocking and Lord ranged from 7% to 13%.

In summary, the results from the differential growth condition were similar to those from the consistent growth condition, with FCIP-1 producing the most accurate classification at the low end of the performance distribution (resulting from the negative bias in the estimates of proficiency), and Stocking and Lord producing the most accurate classification at the high end of the performance distribution (resulting from the positive bias in the estimates of proficiency).

Summary of Results for Study 1

The results of the decision accuracy clearly showed that the FCIP-2 and Stocking and Lord methods produced more accurate results than the FCIP-1 method in both growth conditions. The differences ranged from 2 to 4% depending on the year, which would translate to a difference in accuracy of classification for somewhere between 1400 and 2800 students. An increase in this level of accuracy is non-trivial. Looking at the conditional results, the effect of the underestimation of the performance of FCIP-1 became obvious in the classification of students in the Warning category. FCIP-1 does produce more accurate classifications of students in the Warning category, however, this is not surprising since *all* students are getting an estimate of performance lower than they should, which means that the majority of Warning students will be further below the cut-

score than using the other methods. Because of this, when the results at the higher performance categories are examined, FCIP-1 does not do a very good job of classifying examinees into those categories, relative to the other methods. This is again due to the suppression of the estimates of examinee performance. In contrast, the Stocking and Lord method is superior at the higher levels of performance. Furthermore, when the proficient category is considered, as this is arguably the most important for No Child Left Behind purposes, there is a difference of about 3% of correct classifications between FCIP-1 and either the Stocking and Lord or FCIP-2 methods. A 3% difference in classification accuracy translates into a more accurate classification of about 2100 examinees, which is non-trivial.

IV. Study 2: A Simulation Study to Evaluate the Effect of Changing Equating Methods

The second study was an attempt to understand the effect of changing equating methods after one method has been used for several test administrations. The conditions of this study were identical to Study 1, with the exception that only FCIP-1 was used for equating the first three administrations, and then the FCIP-2 method was used to equate the third and fourth administrations. Again, the accuracy in capturing the changes in the performance as well as the accuracy in the classification of students was examined after the change had been made.

Results

Determination of the Amount of Growth Captured

The results for the error in the growth estimates for both the consistent and differential growth conditions are presented next in Table 15.

Table 15
Error in Growth Estimates for All Methods

		Examinee Growth Conditions	
Year	Equating Method	Consistent	Differential
Year 2	FCIP-1	-0.08	-0.08
Year 3	FCIP-1	-0.04	-0.09
Year 4	FCIP-2	0.00	0.03

The results for the error in the growth estimates showed that when changing from FCIP-1 to FCIP-2 in Year 4, the estimate of growth was very accurate in the consistent growth condition, and slightly overestimated in the differential growth condition. The growth had previously been substantially underestimated, and changing methods corrected that underestimate in the fourth year, but slightly over-corrected in the case of differential growth.

Overall Classification Accuracy Results

The table below presents the overall classification accuracy in Year 4 when the change was made from the FCIP-1 to FCIP-2 method. Both the results for FCIP-1 (no change in method) and FCIP-2 (change in method) are presented to provide context for how different the results were if the method were changed or not changed in the fourth year.

Table 16
Percent of Correctly Classified, Over-classified and Under-classified Examinees, Consistent and Differential Growth

Growth Condition	Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Consistent	Year 4	FCIP-1	0.67	0.25	0.08
Consistent	Year 4	FCIP-2	0.73	0.12	0.15
Differential	Year 4	FCIP-1	0.68	0.12	0.21
Differential	Year 4	FCIP-2	0.70	0.12	0.18

As can be seen in the table, when the change was made, there was an improvement in classification overall for both growth conditions. The improvements for the consistent growth condition were larger (6%) than in the differential growth condition (2%). For the consistent and differential growth conditions, the misclassifications became smaller in number and more balanced when the change was made to the FCIP-2 method.

Conditional Classification Accuracy Results

The conditional classification results are presented below in Tables 17 for consistent growth and Table 18 for differential growth. Again, results for Year 4 are presented for both FCIP-1, which would indicate no change in method, and FCIP-2, which would indicate a change in method.

Table 17
Percent of Correctly Classified, Over-classified, and Under-classified Examinees by True Classification: Consistent Growth

True Classification	Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Warning	Year 4	FCIP-1	0.82	0.17	X
	Year 4	FCIP-2	0.82	0.18	X
Needs Improvement	Year 4	FCIP-1	0.64	0.20	0.17
	Year 4	FCIP-2	0.63	0.22	0.15
Proficient	Year 4	FCIP-1	0.66	0.16	0.17
	Year 4	FCIP-2	0.68	0.17	0.15
Advanced	Year 4	FCIP-1	0.78	X	0.23
	Year 4	FCIP-2	0.80	X	0.20

In the consistent growth condition, in all categories, except Needs Improvement, there was an increase in the percent of correctly classified examinees when the change was made from the FCIP-1 to FCIP-2 method. The change was not dramatic, ranging from 0% to 2%. The types of misclassifications stayed relatively unchanged.

Table 18
Percent of Correctly Classified, Over-classified, and Under-classified Examinees by True Classification: Differential Growth

True Classification	Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Warning	Year 4	FCIP-1	0.79	0.21	X
	Year 4	FCIP-2	0.77	0.23	X
Needs Improvement	Year 4	FCIP-1	0.59	0.20	0.21
	Year 4	FCIP-2	0.61	0.20	0.19
Proficient	Year 4	FCIP-1	0.65	0.14	0.21
	Year 4	FCIP-2	0.68	0.15	0.17
Advanced	Year 4	FCIP-1	0.74	X	0.26
	Year 4	FCIP-2	0.77	X	0.20

In the differential growth condition, the trend was similar, with improvements in the percent of students correctly classified when changing from the FCIP-1 to FCIP2 method, in all categories except Warning. The improvements were about 3% again, and the direction of the misclassification was generally the same with either method.

Summary of Results from Study 2

Study 2 was designed to investigate the effect of changing methods from FCIP-1 to FCIP-2. The results of the study indicated that when the change was made, generally, there was less error in the estimate of growth and the classification of examinees. For the estimation of growth, the improvement was dramatic for the consistent growth condition,

with the growth being estimated perfectly. However, in the case of the differential growth there was a slight overestimate of the amount of growth. For classification accuracy, the results were practically significant. It is important to realize that 3% of 70,000 is not a trivial number of examinees (2100). As such, the change would result in an increase of about 2,100 examinees being properly classified.

V. Study 3: A Simulation Study to Investigate the Effect of Test Difficulty on Equating Accuracy

The third study was conducted to understand how the relative difficulty of the test in relation to the anchor affects the quality of the equating methods in the context of changes in examinee performance. As examinee performance improves, the items on the test may be selected to better match the examinee distribution, however, the anchor items will, by necessity, stay the same across two given administrations. Therefore, the average difficulty of the scoring items and the average difficulty of the anchor items may not be the same. It is a hallmark of equating designs that the anchor test be a miniature version of the whole test, including the relative difficulty of the anchor and scoring tests. Therefore, by having the scoring items and anchor items have different average difficulty parameters, the quality of the equating may be compromised. In this study, two scenarios were considered: (1) Case A: the anchor items are easier than the scoring items (on average) and (2) Case B: the anchor items are more difficult than the scoring items (on average). It should be noted that Case A is the far more realistic situation of the two in Massachusetts. It would be highly inappropriate to construct an easier test each year as the proficiency distribution itself is increasing but this was carried out in Case B as it did give us the opportunity to investigate situations where the operational test was either easier or more difficult than the anchor test.

As in Studies 1 and 2, four administrations of the assessment were simulated. However, this time, the forms were not constructed to be parallel, but rather for the overall difficulty of the test to change across administrations. In the first case (Case A), the scoring items became more difficult by incrementing the b-parameters by .15 each year, while leaving the anchor items the same as in the first study. In the second case (Case B), the scoring tests were easier than the anchor items, although consistent across the administrations. The descriptive statistics for the item parameters are presented in Tables 19 and 20 below.

Table 19
Case A: Mean Parameters for Each Administration by Item Type

Type	Parameter	Year 2	Year 3	Year 4
Scoring: Dichotomous	a	0.98	0.98	0.98
	b	-0.19	-0.04	0.11
	c	0.16	0.16	0.16
Anchor: Dichotomous	a	0.97	0.97	0.97
	b	-0.35	-0.35	-0.35
	c	0.16	0.16	0.16
Scoring: Polytomous	a	1.12	1.12	1.12
	b	0.03	0.18	0.33
Anchor: Polytomous	a	0.93	0.93	0.93
	b	-0.12	-0.12	-0.12

able 20: Case B: Mean Parameters for Each Administration by Item Type

Type	Parameter	Year 2	Year 3	Year 4
Scoring: Dichotomous	a	0.98	0.98	0.98
	b	-0.49	-0.64	-0.79
	c	0.16	0.16	0.16
Anchor: Dichotomous	a	0.98	0.93	0.97
	b	-0.34	-0.35	-0.35
	c	0.16	0.14	0.15
Scoring: Polytomous	a	1.12	1.12	1.12
	b	-.27	-.42	-0.57
Anchor: Polytomous	a	0.93	0.93	0.93
	b	-0.12	-0.12	-0.12

The performance parameters were the same as those used in the initial study, with two different types of growth simulated: consistent or differential. Again, the administrations of the forms were equated using FCIP-1 and FCIP-2 methods as well as the Stocking and Lord method. Simulated examinees were classified in each case using both their true performance parameters and their estimated performance parameters and the number of simulated examinees in each of the categories was tabulated. In addition, the amount of growth that was estimated between administrations was compared to the true amount of growth simulated for each administration and the difference between these quantities was computed for each method, and averaged over replications.

Results

Determination of the Amount of Growth Captured

The results for the error in the growth estimates for both the consistent and differential growth conditions are presented next in Table 21 for the consistent growth condition, and in Table 22 for the differential growth condition.

Table 21
Error in the Estimate of Growth for All Equating Methods, Consistent Growth

		Anchor Item Difficulty Conditions	
Year	Equating Method	Anchor Easier: Case A	Anchor More Difficult: Case B
Year 2	SL	0.04	0.01
	FCIP-1	-0.22	-0.10
	FCIP-2	-0.01	-0.05
Year 3	SL	0.03	0.03
	FCIP-1	-0.15	-0.19
	FCIP-2	0.00	-0.06
Year 4	SL	0.02	0.02
	FCIP-1	-0.15	-0.18
	FCIP-2	-0.03	-0.09

As indicated in the table, in the case of consistent growth, for Case A, where the anchor items were generally easier than the scoring items, the amount of growth was underestimated for the FCIP methods, and overestimated with the Stocking and Lord method. FCIP-1 model produced the least accurate estimate of the growth in almost every case, while FCIP-2 method produced the most accurate estimates. When the anchor items were more difficult than the scoring items, The Stocking and Lord method was the most accurate, although led to a small overestimation of growth. Both the FCIP-1 and FCIP-2 methods led to underestimation of growth, with FCIP-1 producing a large amount of negative bias.

Table 22
Error in the Estimate of Growth for All Methods, Differential Growth

		Anchor Item Difficulty Conditions	
Year	Equating Method	Anchor Easier: Case A	Anchor More Difficult: Case B
Year 2	SL	0.05	0.02
	FCIP-1	-0.22	-0.11
	FCIP-2	-0.01	-0.05
Year 3	SL	-0.05	-0.06
	FCIP-1	-0.21	-0.15
	FCIP-2	-0.06	-0.07
Year 4	SL	-0.06	-0.06
	FCIP-1	-0.22	-0.18
	FCIP-2	-0.08	-0.08

In the case of differential growth, all methods tended to underestimate the growth regardless of the relative difficulty of the anchor and scoring items. FCIP-1 method produced the least accurate estimate of the growth while FCIP-2 and Stocking and Lord methods performed similarly.

Overall Classification Accuracy Results

The overall classification accuracy results are presented below in Tables 23 and 24. Table 23 provides the results for the consistent growth conditions for Case A, while Table 24 presents the results for the consistent growth condition for Case B.

Table 23
Percent of Correctly Classified, Over-classified, and Under-classified Examinees, Overall, Consistent Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.71	0.15	0.13
	FCIP-1	0.65	0.06	0.30
	FCIP-2	0.71	0.13	0.16
Year 3	SL	0.72	0.16	0.12
	FCIP-1	0.57	0.03	0.41
	FCIP-2	0.72	0.13	0.16
Year 4	SL	0.73	0.16	0.11
	FCIP-1	0.47	0.01	0.52
	FCIP-2	0.73	0.11	0.17

The pattern of results for Case A, where the anchor items were easier than the scoring items is similar to that observed in Study 1, where the scoring items and anchor items were of similar difficulty. Stocking and Lord and FCIP-2 methods performed identically in terms of the percent of correct classifications while FCIP-1 led to a greater number of misclassifications, at about 6%-26% difference between FCIP-1 and the other two methods. However, for FCIP-1, the rate of misclassifications for Case A was much higher than in Study 1, indicating that when the anchor items were easier than the scoring items, FCIP-1 did not perform as well as when the anchor and scoring items were of equal difficulty. Again, the FCIP methods both produced more under-classifications than over-classifications, while Stocking and Lord produced more over-classifications.

Table 24
Percent of Correctly Classified, Over-classified, and Under-classified Examinees, Overall, Consistent Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.69	0.15	0.16
	FCIP-1	0.55	0.03	0.42
	FCIP-2	0.69	0.14	0.17
Year 3	SL	0.69	0.17	0.14
	FCIP-1	0.49	0.02	0.49
	FCIP-2	0.69	0.14	0.18
Year 4	SL	0.70	0.17	0.13
	FCIP-1	0.44	0.01	0.55
	FCIP-2	0.69	0.13	0.18

In Case B, a similar pattern arose: FCIP-2 and Stocking and Lord methods led to similar overall correct classifications, while the FCIP-1 method led to less accurate classifications than either of the other two methods. Again, the Stocking and Lord method tended to over-classify examinees, while the FCIP-2 method tended to under-classify examinees. As seen previously, the FCIP-1 method tended to under-classify students in this case as well.

The same results for the differential growth condition are presented next in Tables 25 to 26. Table 25 presents the Case A results, where the anchor items are easier than the scoring items and Table 26 presents the Case B results, where the anchor items are more difficult than the scoring items.

Table 25
Percent of Correctly Classified, Over-classified, and Under-classified Examinees, Overall, Differential Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.70	0.16	0.14
	FCIP-1	0.64	0.06	0.30
	FCIP-2	0.70	0.14	0.13
Year 3	SL	0.70	0.17	0.13
	FCIP-1	0.54	0.03	0.43
	FCIP-2	0.70	0.14	0.16
Year 4	SL	0.71	0.16	0.13
	FCIP-1	0.42	0.01	0.57
	FCIP-2	0.70	0.12	0.18
	FCIP-1 to 2	0.70	0.12	0.18

The results for Case A were quite similar to those in Study 1 for Stocking and Lord and FCIP-2. However, as before, FCIP-1 performed notably worse in this case, and led to more under-classifications than in the previous case.

Table 26
Percent of Correctly Classified, Over-classified, and Under-classified Examinees, Overall, Differential Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.69	0.15	0.16
Year 2	SL	0.68	0.16	0.16
	FCIP-1	0.58	0.04	0.38
	FCIP-2	0.68	0.15	0.18
Year 3	SL	0.67	0.17	0.16
	FCIP-1	0.57	0.04	0.39
	FCIP-2	0.67	0.15	0.19
Year 4	SL	0.67	0.17	0.16
	FCIP-1	0.57	0.04	0.39
	FCIP-2	0.66	0.14	0.20

Again, similar to the consistent growth case, Stocking and Lord and FCIP-2 methods led to similar accuracy in classification, while the FCIP-1 method led to less accurate results for Case B. Again, the Stocking and Lord method tended to over-classify students, and the FCIP methods tended to under-classify students.

Conditional Classification Accuracy Results

The results were then broken down into separate results for examinees in each performance category. Tables 27 to 30 are the results for Case A where the anchor items are easier than the scoring items, and Tables 31 to 34 are the results for Case B where the anchor items are more difficult than the scoring items. Note too that in case A, the scoring items better match the proficiency distribution. In case B, the match got worse over time. Correspondingly proficiency estimates were better in Case A than in Case B.

Table 27
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Consistent Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.81	0.19	X
	FCIP-1	0.94	0.06	X
	FCIP-2	0.83	0.17	X
Year 3	SL	0.79	0.21	X
	FCIP-1	0.99	0.01	X
	FCIP-2	0.83	0.17	X
Year 4	SL	0.73	0.13	X
	FCIP-1	0.97	0.03	X
	FCIP-2	0.80	0.20	X

For the examinees whose true classification was Warning, the results were similar to those found in Study 1, where the FCIP-1 method produced the most accurate classifications, and the Stocking and Lord method produced the least accurate results. The differences were large, ranging from 13% to 20%.

Table 28
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Consistent Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.62	0.24	0.14
	FCIP-1	0.59	0.09	0.32
	FCIP-2	0.63	0.21	0.16
Year 3	SL	0.60	0.26	0.14
	FCIP-1	0.50	0.04	0.46
	FCIP-2	0.62	0.21	0.27
Year 4	SL	0.61	0.27	0.22
	FCIP-1	0.40	0.02	0.58
	FCIP-2	0.63	0.21	0.16

In the case where the true classification was Needs Improvement, FCIP-1 led to the fewest students being correctly classified, and Stocking and Lord and FCIP-2 methods produced similar results. The differences were large again, ranging from 3% to 21%. For Stocking and Lord and FCIP-2 methods the misclassifications were more likely to be over-classifications, while for the FCIP-1 method, the misclassifications tended to be under-classifications.

Table 29
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Consistent Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.67	0.17	0.16
	FCIP-1	0.60	0.06	0.34
	FCIP-2	0.68	0.14	0.18
Year 3	SL	0.67	0.21	0.12
	FCIP-1	0.53	0.04	0.43
	FCIP-2	0.69	0.16	0.15
Year 4	SL	0.65	0.23	0.12
	FCIP-1	0.42	0.02	0.56
	FCIP-2	0.68	0.15	0.17

For the examinees with a true classification of Proficient, the differences between the methods remained dramatic in some cases, with the FCIP-1 method producing the least accurate results and FCIP-2 method the most accurate results. The differences ranged from 8% to 26%. Stocking and Lord produced results very similar to FCIP-2 with differences between these two methods ranging from 1% to 3%. The Stocking and Lord method continued to produce more over-classifications than under-classifications, while the FCIP-2 method produced relatively balanced errors. FCIP-1 method produced more under-classifications than over-classifications.

Table 30
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Consistent Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.81	X	0.19
	FCIP-1	0.60	X	0.40
	FCIP-2	0.77	X	0.23
Year 3	SL	0.83	X	0.17
	FCIP-1	0.50	X	0.50
	FCIP-2	0.77	X	0.23
Year 4	SL	0.88	X	0.12
	FCIP-1	0.44	X	0.56
	FCIP-2	0.80	X	0.20

In the case where the true classification of the examinees was Advanced, Stocking and Lord method produced the most accurate results, while the FCIP-1 method produced the least accurate results. The differences were large, ranging from 21% to 44%. FCIP-2 and Stocking and Lord methods were more similar, with differences ranging from 4% to 8%.

Tables 31 to 34 present the same results for the case of differential growth.

Table 31
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Differential Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.75	0.25	X
	FCIP-1	0.91	0.09	X
	FCIP-2	0.78	0.22	X
Year 3	SL	0.69	0.31	X
	FCIP-1	0.95	0.05	X
	FCIP-2	0.76	0.14	X
Year 4	SL	0.67	0.33	X
	FCIP-1	0.98	0.02	X
	FCIP-2	0.76	0.24	X

For examinees whose true classification was Warning, the equating was most accurate when the FCIP-1 method was used. This result was similar to the result found in Study 1, and also in Case A, with consistent growth. The difference between the FCIP-1 method and the other two methods was sizable, and the difference in correct classifications between FCIP-1 and Stocking and Lord, which performed the worst, ranged from 16% to 31%. The FCIP-2 method did slightly better than the Stocking and Lord method, but still was quite different from the FCIP-1 method.

Table 32
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Differential Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.62	0.24	0.14
	FCIP-1	0.58	0.08	0.34
	FCIP-2	0.62	0.21	0.17
Year 3	SL	0.62	0.27	0.11
	FCIP-1	0.50	0.05	0.45
	FCIP-2	0.62	0.22	0.16
Year 4	SL	0.61	0.27	0.12
	FCIP-1	0.37	0.02	0.61
	FCIP-2	0.61	0.20	0.19

In the Needs Improvement category, the differences between the methods were not as great as in the Failing category, and the Stocking and Lord and FCIP-2 methods led to identical results. FCIP-1 method had the fewest examinees classified correctly, and the difference between FCIP-1 and the other two methods, in terms of correct classification, ranged from 4% to 24%. The misclassifications for the Stocking and Lord method were generally over-classifications. For the FCIP-2 method, the errors were more balanced, with a slightly higher rate of over-classification relative to under-classification. For the FCIP-1 method, nearly all of the misclassifications were under-classification errors.

Table 33
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Differential Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.68	0.18	0.14
	FCIP-1	0.60	0.07	0.33
	FCIP-2	0.68	0.15	0.17
Year 3	SL	0.69	0.18	0.13
	FCIP-1	0.50	0.04	0.56
	FCIP-2	0.69	0.15	0.16
Year 4	SL	0.69	0.18	0.13
	FCIP-1	0.38	0.01	0.61
	FCIP-2	0.68	0.14	0.17

For the case of Proficient examinees, Stocking and Lord and FCIP-2 methods performed almost identically, with the exception of Year 4, where there was a 1% difference in correct classifications between the two, with the Stocking and Lord method being slightly more accurate. FCIP-1 method had a substantially worse rate of correct classification, especially by Year 4. The difference between the FCIP-1 method and the other methods ranged from 8% in Year 2 to 31% in Year 4. Again, the FCIP-1 method had a vast majority of its misclassifications as under-classifications, while Stocking and Lord method had more over-classifications than under-classifications and the FCIP-2 method had slightly more under-classifications than over-classifications.

Table 34
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Differential Growth, Case A

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.79	X	0.21
	FCIP-1	0.58	X	0.42
	FCIP-2	0.76	X	0.24
Year 3	SL	0.80	X	0.20
	FCIP-1	0.47	X	0.53
	FCIP-2	0.76	X	0.24
Year 4	SL	0.82	X	0.18
	FCIP-1	0.37	X	0.63
	FCIP-2	0.78	X	0.22

As with the consistent growth condition, and the results of Study 1, for students with a true classification of Advanced, Stocking and Lord method had the highest rate of correct classifications, and the FCIP-1 method had the lowest rate. FCIP-2 method was similar to the Stocking and Lord method, although slightly less accurate. The differences between FCIP-1 and Stocking and Lord methods were sizable, ranging from 21% to 45%.

Tables 35 to 38 present the conditional classification results for Case B, with consistent growth, where the anchor items are more difficult than the scoring items.

Table 35
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Consistent Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.80	0.20	X
	FCIP-1	0.70	0.30	X
	FCIP-2	0.82	0.18	X
Year 3	SL	0.80	0.20	X
	FCIP-1	0.65	0.35	X
	FCIP-2	0.83	0.17	X
Year 4	SL	0.72	0.28	X
	FCIP-1	0.52	0.48	X
	FCIP-2	0.77	0.23	X

In the case where the true examinee classification was Warning, the FCIP-2 method produced the most accurate classifications. This result is different from the results in the other cases, where FCIP-1 method produced the most accurate results. In this case, where the anchor items are more difficult than the scoring items, however, there was a marked decline in the performance of the FCIP-1 method. Stocking and Lord and FCIP-2 methods performed similarly.

Table 36
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Consistent Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.61	0.25	0.14
	FCIP-1	0.58	0.15	0.27
	FCIP-2	0.61	0.24	0.15
Year 3	SL	0.59	0.29	0.12
	FCIP-1	0.50	0.18	0.32
	FCIP-2	0.60	0.25	0.15
Year 4	SL	0.59	0.30	0.11
	FCIP-1	0.45	0.17	0.38
	FCIP-2	0.62	0.14	0.14

For the Needs Improvement category, the FCIP-2 and Stocking and Lord methods performed similarly. In Year 2, all three methods led to identical results for the percent of correct classification. Across years, however, there was a decline in the FCIP-1 method, with correct classification changing 13% across the administrations. There was also a slight decline in the accuracy of classification using the Stocking and Lord method, with a change of 2% across administrations. For misclassifications, the FCIP-2 and Stocking and Lord methods had predominantly over-classifications relative to under-classifications, while FCIP-1 continued to produce larger under-classifications.

Table 37
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Consistent Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.66	0.17	0.17
	FCIP-1	0.60	0.14	0.26
	FCIP-2	0.66	0.15	0.19
Year 3	SL	0.65	0.21	0.14
	FCIP-1	0.61	0.16	0.33
	FCIP-2	0.67	0.17	0.14
Year 4	SL	0.63	0.24	0.13
	FCIP-1	0.55	0.15	0.30
	FCIP-2	0.66	0.17	0.17

When the true classification of examinees was Proficient, again, the results of the FCIP-2 and Stocking and Lord methods were quite similar, while FCIP-1 produced less accurate results than either of the other two methods. Again, in general, the misclassifications tended to be over-classifications for the FCIP-2 and Stocking and Lord methods, while FCIP-1 produced more under-classifications.

Table 38
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Consistent Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.75	X	0.25
	FCIP-1	0.68	X	0.32
	FCIP-2	0.72	X	0.28
Year 3	SL	0.77	X	0.23
	FCIP-1	0.66	X	0.34
	FCIP-2	0.72	X	0.28
Year 4	SL	0.82	X	0.18
	FCIP-1	0.66	X	0.34
	FCIP-2	0.75	X	0.25

The case where the true classification of the examinees was Advanced led to results that were fairly similar across the FCIP-2 and Stocking and Lord methods, while the accuracy of the FCIP-1 method was dramatically worse.

The same results for differential growth are presented next in Tables 39 to 42.

Table 39
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Warning, Differential Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.81	0.19	X
Year 2	SL	0.73	0.27	X
	FCIP-1	0.71	0.29	X
	FCIP-2	0.76	0.24	X
Year 3	SL	0.69	0.31	X
	FCIP-1	0.61	0.39	X
	FCIP-2	0.74	0.26	X
Year 4	SL	0.66	0.34	X
	FCIP-1	0.50	0.50	X
	FCIP-2	0.74	0.26	X

In the case where the true classification of the examinees was Warning, again there was a trend that FCIP-2 produced the most accurate results, and FCIP-1 the least accurate results. The differences between the methods got larger over time, ranging from 5% in Year 2 to 26% in Year 4. Stocking and Lord produced results in between those of the two FCIP methods.

Table 40
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Needs Improvement, Differential Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.62	0.22	0.16
Year 2	SL	0.61	0.25	0.14
	FCIP-1	0.59	0.14	0.27
	FCIP-2	0.61	0.23	0.16
Year 3	SL	0.59	0.30	0.11
	FCIP-1	0.51	0.18	0.33
	FCIP-2	0.60	0.25	0.15
Year 4	SL	0.59	0.30	0.11
	FCIP-1	0.46	0.18	0.39
	FCIP-2	0.59	0.24	0.17

In the case of Needs Improvement, Stocking and Lord and FCIP-2 methods produced almost identical results, with the FCIP-1 method producing slightly less accurate results than the other two methods. Similar to the consistent growth conditions, the FCIP-2 and Stocking and Lord methods produced over-classifications, while the FCIP-1 method produced under-classifications.

Table 41
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Proficient, Differential Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.67	0.15	0.18
Year 2	SL	0.68	0.17	0.15
	FCIP-1	0.60	0.14	0.26
	FCIP-2	0.67	0.15	0.18
Year 3	SL	0.68	0.17	0.15
	FCIP-1	0.61	0.16	0.33
	FCIP-2	0.66	0.16	0.18
Year 4	SL	0.67	0.18	0.15
	FCIP-1	0.55	0.15	0.30
	FCIP-2	0.66	0.15	0.19

For Proficient examinees, again, the FCIP-2 and Stocking and Lord methods performed very similarly, while the FCIP-1 method was substantially less accurate. The Stocking and Lord method was consistently better than the two FCIP methods, however the differences were small, and ranged from 1% to 3% for correct classifications. The misclassification errors for Stocking and Lord were slightly more likely to be over-estimates, the FCIP-2 method had more balanced errors, with a slight tendency toward under-classification, while FCIP-1 consistently led to under-classification.

Table 42
Percent of Correctly Classified, Over-classified, and Under-classified Examinees; True Classification: Advanced, Differential Growth, Case B

Year	Equating Method	Correctly Classified	Over Classified	Under Classified
Year 1		0.71	X	0.28
Year 2	SL	0.72	X	0.28
	FCIP-1	0.68	X	0.32
	FCIP-2	0.70	X	0.30
Year 3	SL	0.73	X	0.27
	FCIP-1	0.66	X	0.34
	FCIP-2	0.70	X	0.30
Year 4	SL	0.74	X	0.26
	FCIP-1	0.66	X	0.34
	FCIP-2	0.70	X	0.30

In the case where the true classification of examinees was Advanced, the FCIP-2 and Stocking and Lord methods had a higher rate of correct classification than FCIP-1, and were similar to each other. Stocking and Lord method was consistently better than the FCIP-2 method. The differences among the two methods were not large, while the differences between either FCIP-2 or Stocking and Lord and FCIP-1 were large.

Summary of Results for Study 3

For the overall results, in terms of capturing the change in the performance of examinees from administration to administration, the results were clear. Both the FCIP-2 and Stocking and Lord methods produced similar results, and were more accurate in capturing growth than the FCIP-1 method. FCIP-1 consistently led to underestimates of growth, while the other methods varied depending upon the relative difficulty of the anchor and scoring items.

For overall classification rates, as well as conditional classification rates, Stocking and Lord and FCIP-2 methods produced similar results, and had greater accuracy of classification than the FCIP-1 method. FCIP-1 consistently produced under-classifications as opposed to over-classifications. This is not surprising given the underestimate of growth.

VI. Summary of Results from the Three Studies

Results from Study 1 indicated that the FCIP-2 and the Stocking and Lord methods produced more accurate results than the FCIP-1 method (1) in assessing growth (both consistent and differential) and (2) in assigning examinees correctly to performance categories. Regarding the latter, the difference ranged from 2 to 4% depending on the year, which translated to a difference in accuracy of classification for somewhere between 1,400 and 2,800 examinees (with a total of 70,000 examinees in the sample). When the proficient category is considered, as this is arguably the most important for No Child Left Behind purposes, there is a difference of about 3% of correct classifications between the FCIP-1 method and either the Stocking and Lord or the FCIP-2 method. A 3% difference in classification accuracy translates into more accurate classification for about 2,100 examinees. This is a practically significant difference and could easily be used to justify either of the better equating methods.

Study 2 was designed to investigate the effect of changing methods from FCIP-1 to FCIP-2. The results of the study indicated that when the change was made, generally, there was less error in the estimate of growth and the classification of examinees. The improvement was even more substantial for the consistent growth condition than for the differential growth condition. The difference in classification accuracy was about 3%, which again translated to about 2,100 examinees in the state. As such, the change would result in an increase of about 2,100 examinees being correctly classified.

The results of Study 3 indicated that in the case where the anchor items were easier than the scoring items (Case A), the FCIP-2 method produced the best estimate of the true growth. For overall classification rates, the Stocking and Lord and FCIP-2 methods produced similar results, and both produced greater accuracy of examinee classification than the FCIP-1 method. These are important findings because Case A is a common scenario in practice—it is the case where as examinees demonstrate more proficiency over time, test difficulty is revised accordingly to improve the precision with which examinee proficiency is estimated. In this case too, the difficulty of the equating items remains about the same from year to year. Findings from Case B were consistent with those of Case A, where FCIP-1 led to dramatically worse estimates of growth and classification of examinees than either FCIP-2 or Stocking and Lord.

The overall trends in the findings in these three studies indicate that the FCIP-2 and Stocking and Lord methods performed similarly to each other, and produced more accurate estimates of growth, and more accurate classification of examinees than the FCIP-1 method. The differences do not seem great (typically between 1 and 4%), but they are of considerable practical consequence, and the FCIP-2 method could be implemented with only a minor change or two in the current equating process in Massachusetts. In terms of changing methods from FCIP-1 to FCIP-2, the results of Study 2 indicated that when the change was made, there was a reduction in the error of the estimation of growth and an improvement in the classification accuracy of examinees. While the changes may not seem dramatic, due to the large sample sizes in the MCAS

program, as noted previously, differences of even 3% can affect the performance classifications of a large number of examinees—about 2,100 examinees.

Therefore, taken together, the overall results of these three studies indicate that using either FCIP-2 or Stocking and Lord equating methods appear to lead to somewhat more accuracy than using the FCIP-1 method. As the simulations (with the exception of Case B in Study 3) very much match the tests and the conditions in the MCAS program from year to year, generalization of the main findings to other grades 3 to 8 ELA and mathematics tests would seem to be warranted. The results from the research based on theoretical support (it is best to reset the prior proficiency distribution after each stage in item parameter estimation to reflect the best available estimates of the mean and standard deviation of the proficiency scores) as well as the empirical evidence from the three studies, strongly support a decision to switch from the FCIP-1 to the FCIP-2 equating method.

References

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71*, 229–250.

Keller, L. A., Keller, R. R., & Baldwin, S. G. (2007, April). *The effect of changing equating methods on monitoring growth in mixed-format tests.* Paper presented at the meeting of the National Council on Measurement in Education. Chicago, IL.

Keller, L. A., & Keller, R. R. (2007a). *A comparison of transformation methods and calibration methods on the classification of students over time* (Center for Educational Assessment Research Report No. 664). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Keller, L. A., & Keller, R. R. (2007b). *The effect of changing equating methods on the classification of students in mixed-format tests* (Center for Educational Assessment Research Report No. 665). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*(4), 355-81.

Li, Y. H., Griffith, W. D., & Tam, H. P. (1997). *Equating multiple tests via an IRT linking design: Utilizing a single set of anchor items with fixed common item parameters during the calibration process.* Paper presented at the meeting of the Psychometric Society, Knoxville, TN.

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18*, 199–215.