Adjusting for Year to Year Rater Variation in IRT Linking –An Empirical Evaluation

Shu Jing Yen
Charles Ochieng
Hillary Michaels
Greg Friedman

CTB/McGraw-Hill

Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Canada, 2005.

Abstract

The main purpose of this study was to examine a polytomous IRT-based linking procedure that adjusts for rater variations. Test scores from two administrations of a statewide reading assessment were used. An anchor set of Year 1 students' constructed responses were rescored by Year 2 raters. To adjust for year-to-year rater variation in IRT linking, a two-step approach was used. First, the Year 1 item parameters for the constructed response items were re-estimated using the ratings assigned by the Year 2 raters. Through this recalibration process, the Year 1 item parameters for the constructed response items were adjusted for rater variation. Second, the Stocking-Lord equating method was used to place the Year 2 form on the Year 1 scale using all the common items between forms.

This method was compared with two alternative methods: traditional IRT linking study that links the test forms using a) all the common items without the rater adjustment and b) the common selected response items. The success of the Stocking-Lord procedure was compared among the three methods. The differences in the test characteristic curves and the students' scale score distributions produced by the three methods were also compared.

Significant shifts in the parameters after rater adjustment were found for one (grade 8) of the three grades examined. The p-values and TCCs shifted across years when adjusted for rater effects. The impact of the parameter shifts and TCCs manifested in the changes in the proficiency classification before and after adjustment. However, a systematic bias might have been introduced in the equating process while trying to adjust for the rater variation through the equating process. Further studies are needed to address this problem in greater detail. Using only the common SR items seems to produce satisfactory results. Thus, in the case where it is not feasible to integrate rater adjustment in the equating process, using SR item anchors is a better approach than using the mixed-item anchors without adjusting for rater effect.

The results of the study suggest that raters were not consistently more severe or more lenient between grades, but the resulting rater error (severity or leniency) affected the scores and thereby produced misleading results if not taken into account.

Introduction

Constructed response items have become a standard part of educational assessment. The inclusion of these items is motivated primarily by validity arguments. They are thought to be a more authentic evaluation of student's competence.. Several critical measurement issues related to the use of mixed item format tests have been investigated by measurement experts, e.g., Tate (1999, 2000); Wilson & Wang (1995). One of these issues in linking mixed item format tests is score comparability across test administration years. One common equating design used in linking or equating tests from year to year is item response theory (IRT) scaling using a non-equivalent, common item equating design. The traditional linking method often applied to linking test forms administered in different years is an anchor test design. If the constructed response (CR) items are included as part of the anchor set with selected response (SR) items, there may be the presence of year-to-year rater variation. The need to have the anchor item set include all elements of the test specifications suggests such inclusion. However, year-to-year rater variation might result in changes in the parameters of the CR items, making such inclusion inappropriate.

Research on rater issues has shown that rater severity has significant impact on student's scores (Engelhard, 1992, Lunz, Writing, & Linacre, 1990, Patz, 1999) and therefore cannot be ignored. Wilson and Wang (1995) found that although inter-rater correlations were moderately high in their study, the range of rater severities was quite large, and that the rater effects on the constructed response items can impact student scores. Studies (Ito & Sykes, 1998) have shown that even if a test has a small number of constructed response items, such as CRs only contributing about 30% of total points

possible, a substantial amount of year to year rater fluctuation can yield relatively large score differences even after the test is equated through selected response items to an existing scale. Fitzpatrick et al (1998) found that the variation of rater severity changed every year for a statewide assessment. And that these variations could affect students' proficiency classifications, indicating the need to adjust for rater effects during the equating process. Therefore, linking procedures that do not take rater severity into account may produce misleading results.

Tate (1999) argued that raters differ in discrimination and severities over time and that applying the standard linking method through common CR items result in incorrect results since year to year changes in the anchor item parameters may be due to both to changes in the anchor item parameters and to changes in rater discrimination and severity. Although a relatively convenient option is to use only selected response (SR) items in the anchor sets, as the number of constructed response items increases in a test, the impact of rater severity or leniency on scores becomes greater. Tate (2000) has shown that linking based on SR anchor items can sometimes produce accurate results for a mixed item format test. However, such a linking method is only defensible if the selected response items and the constructed response items measure the same construct.

Since the use of only SR items for linking may lead to serious linking bias under some conditions, it is suggested by some authors (Fitzpatrick et al, 1998; Wilson & Wang, 1995; Tate, 1999, 2000) that it is necessary to apply an equating methods to adjust for the rater effects. Tate (1999, 2000) adjusted for rater effects through IRT linking using simulated data. His procedure involves conducting a linking study in which a sample of anchor item papers for examinees from the prior year is inserted into the

scoring process for the current year. He further stipulated that the sample of prior year papers needs to be representative of all the papers and that the specific raters judging the anchor papers from the prior year must be the same raters that judge the current year's papers for that item. An initial estimate of the item parameters were obtained for both years. Then the item parameters associated with the CR anchor items on the prior year papers were re-estimated using the scores assigned by the current year's raters. These 'adjusted' item parameters were then used in the final scaling for the current year form. The final scale for the current year was obtained by conducting a Stocking & Lord (1983) equating procedure using all the common items between the forms. Since the CR item parameters have been adjusted to reflect the rating standards for the current year's raters, cross year rater effect have been held constant through the linking procedure, the year-to-year change in ability was correctly reflected in the students' ability estimates.

Wilson and Wang (1995) used a Random Coefficient Multi-nominal Logit Model to model both the item difficulties and rater severities in the same model. In particular, the item parameters of the constructed response items were decomposed into linear combinations of the item difficulties and rater severities. Three parallel test forms were administered to three different groups of examinees. There was one common SR item across the three forms. Two of the forms have additional SR items in common. Two of the forms have common raters. Common items and common raters were used to link the three forms together. The item parameters of the three forms were estimated simultaneously rather than separately. Rater effect was removed using the simultaneous calibration process.

Both Tate's (1999, 2000) and Wilson and Wang (1995) methods attempt to adjust for rater effect in the IRT linking process. Tate's method is applicable to IRT models in general and most IRT calibration and equating software is sufficient. Wilson and Wang's method is limited to the Rasch family of models and it requires specialized programs. Although Tate's method shows a lot of promises, only the simulation data were used in the study. More research using real test data is needed.

The purpose of this study is to evaluate the success of adjusting cross year rater variation on IRT linking using empirical data. An IRT linking similar to the method proposed by Tate (1999, 2000) was used to adjust for rater variation between years. This study differs from the Tate's studies in several ways. In Tate's simulation studies, using a two-parameter model and graded response model for data calibration, it was assumed that the Year 2 raters rescored a sample of anchor item papers from examinees in Year 1. This study uses real test data from a state assessment using a three-parameter model and the partial credit model. The entire set of anchor item papers from examines in Year 1 was rescored in Year 2. Lastly, Tate assumed that specific raters who judged the prior year's papers for the anchor item were the same raters who judged the current year's papers for that item. This assumption is not likely to hold in live, large-scale testing situations. In this study, the raters that judged the prior year's papers are not the same raters as those who judged the current year's papers. However, training protocols, processes, and materials were the same between the two years.

Method

Data from a statewide reading assessment were used to investigate the effect of cross year rater variation on the year-to-year form equating. Year 1 and Year 2 forms

were constructed to be parallel. The test consisted of 75, 93, and 91 SR items and 12, 14, and 16 CR items for grade 4, 6, and 8, respectively. All the CR items were scored on a three-point scale (0-2). The number of score points and the percentage of the total number of score points on the test attributed to the non-anchor items and the anchor items by item type are presented in Table 1. The anchor item block consists of both the SR and the CR items.

A three-parameter logistic model was applied to the selected response items (Lord, 1980) and two-parameter partial credit model was used for the constructed response items (Muraki, 1992.) Items within each grade were calibrated using PARDUX software program (Burket, 2002.)

Table 1. The Number of Score Points by Item Type

	Non-Anchor		Anchor	
	SR	CR	SR	CR
Reading				
Grade 4	65(75%)	4(5%)	10(11%)	8(9%)
Grade 6	79(74%)	4(4%)	14(13%)	10(9%)
Grade 8	80(75%)	6(6%)	11(10%)	10(9%)

The test was administered to approximately 40,000 students per grade. So that complete blocks of rater data were available, two samples, designated as Year 1 and Year 2 were selected. The Year 1 stratified random sample consisted of about 1,200 students drawn from the population of test takers in the state. Similar sampling procedures were

used to obtain the Year 2 examinees of about 2,000. Each rater scored only a portion of the CR items on a test. This is a frequently adopted strategy for reducing the halo effect.

To evaluate change in the ratings of the populations of raters between years requires comparisons of the ratings obtained for the student papers for the Year 1 tests. Raters that judged the first year papers were not the same raters that judged the second year papers. However, the same training process and materials were used in training the raters in the two consecutive years. Responses to the constructed response items for each student, in the form of their original papers, were rescored by a representative sample of raters who scored the Year 2 operational examination. This sample of readers was retained towards the end of the scoring session to rescore the Year 1 student papers. These raters were not aware of the previous year's score.

Raw-score statistics for the constructed response items are presented in Table 2. The differences in raw score means and standard deviations between years tend to be small. The exceptions are item 99 in grade 6 and items 91, 92, and 93 in grade 8. Item 91 in grade 8 showed the largest raw score mean difference between years. The difference of 0.65 is especially large for a two-point item. The generally small differences in means suggest that the 1st- and 2nd-year raters, on average, awarded very similar scores. However, the year 2 raters tended to be more lenient than year 1 raters in grading the CR items.

The percentage of perfect, adjacent and discrepant agreement between years is presented in Table 3. The percent perfect agreement varied from 40% to 88%. Grade 8 tended to have lower perfect agreement rates compared to grades 4 and 6. Again, item 99 in grade 8 showed a very low agreement rate between the Year 1 and Year 2 raters.

Table 2

Mean and Standard Deviation of Item Scores Assigned by 1st and 2nd Year Raters

Content	Grade	Item	Mean by Year1 Raters (y1)	Mean by Year 2 Raters (y2)	D= (y1)-(y2)	SD (y1)	SD(y2)
Reading	Grade 4	Item 78	1.13	1.04	0.09	0.74	0.72
		Item 79	0.96	0.97	-0.01	0.90	0.91
		Item 80	0.87	0.88	-0.01	0.78	0.84
		Item 81	0.68	0.77	-0.09	0.83	0.82
	Grade 6	Item 85	0.99	1.04	-0.05	0.67	0.69
		Item 93	1.18	1.23	-0.05	0.83	0.79
		Item 94	0.83	0.91	-0.08	0.66	0.76
		Item 99	1.04	0.80	0.24	0.75	0.63
		Item 100	0.88	0.95	-0.07	0.81	0.83
	Grade 8	Item 84	1.27	1.20	0.07	0.72	0.75
		Item 91	0.68	1.33	-0.65	0.65	0.67
		Item 92	0.97	1.46	-0.49	0.79	0.72
		Item 93	1.08	1.27	-0.19	0.83	0.86
		Item 99	1.04	1.02	0.02	0.61	0.64

Table 3
Across-Year Rater agreements

Content	Grade	Item	%_Perfect Agreement	%_Adjacent Agreement	%_Discrepant	Correlation	Weighted Kappa
Reading	Grade 4	Item 78	69	30	1	0.69	0.69
		Item 79	88	11	1	0.91	0.91
		Item 80	67	31	2	0.71	0.71
		Item 81	82	17	1	0.85	0.84
	Grade 6	Item 85	72	27	1	0.68	0.68
		Item 93	83	17	0	0.87	0.87
		Item 94	70	29	1	0.70	0.69
		Item 99	63	36	1	0.66	0.62
		Item 100	62	34	4	0.62	0.62
	Grade 8	Item 84	71	27	2	0.69	0.69
		Item 91	40	54	6	0.60	0.40
		Item 92	53	37	10	0.55	0.46
		Item 93	80	19	1	0.85	0.83
		Item 99	71	28	1	0.62	0.62

Analysis

The polytomous IRT-based linking procedure used in this study to adjust for rater variation in the linking process is summarized before. First, the Year 1 item parameters for the CR items were estimated using the ratings assigned by the Year 2 raters. Second, the Stocking-Lord (Stocking & Lord, 1983) equating method was used to place the Year 2 test on the Year 1 scale using all the common items. Since the year-to-year rater variation has been accounted for in the first step, the changes in the constructed response anchor item parameters are reflective of changes in item difficulty only as opposed to a combination of rater variation and item difficulty. For ease of discussion, this procedure is referred to as mixed-item anchors with rater adjustment.

Two alternative methods were also implemented for comparison purposes. One uses mixed-item anchors with no rater adjustment and the other uses SR items as anchor. The former is a simply a traditional IRT linking procedure using both the SR and the CR items as anchors. The latter uses only the SR items as anchor. In both of these two procedures, item parameters for the CR items were estimated using the ratings assigned by the Year 1 raters. For both of these two procedures, Year 2 item calibration was conducted. Then the Stocking-Lord equating method was used to place the Year 2 test on the Year 1 scale using the appropriate anchors.

The mixed-item anchors with rater adjustment procedure tries to integrate rater adjustment into the equating process. Operationally, Year 1 item parameters for the CR items were re-estimated using the ratings assigned by the Year 2 raters. This was accomplished by replacing the ratings assigned by the Year 1 raters with those of the Year 2 raters for the same students. Note that item parameters for the SR items were

also re-estimated in this process. Then the Stocking-Lord (Stocking & Lord, 1983) equating method was used to place the Year 2 test on the Year 1 scale using all the common items between forms.

Results and Discussions

Summary statistics for equating Year 2 and Year 1 form are shown from Table 4 to Table 6 for each grade. Table 4 showed that grade 4 students who were administered Year 2 form had a number correct score mean of 9.78 on the common SR and CR items. Students who were administered Year 1 form had a number correct score mean of 9.95 and 9.98 (after rescore) on the same items. Thus, based on the combined SR and CR score, Year 2 students appear to be *lower* performing students than the Year 1 students. Similar conclusion would be made when comparing the number correct score mean between years using either the common SR items or the common CR items. Based on this finding, using mixed-item anchors, mixed-item anchors adjusted for rater variation, or SR item anchors should produce similar equating results.

For grade 6, based on the common SR and CR items, Year 2 students appear to be slightly *lower* performing than the Year 1 students. However, the differences in the number correct score mean between years were very small. Similar conclusion would be drawn when comparing the common CR or the common SR means between years. Therefore, it is expected that Year 2 students' performance will be very similar to those of the Year 1 students. Furthermore, Year 2 students' scores might not be affected by the choice of the anchor set for equating. Using mixed-item anchors, mixed-item anchors adjusted for rater variation, or SR item anchors should produce similar results.

For grade 8, comparing the performance of Year 2 and Year 1 students using the number correct score mean between years were difficult because different conclusions would be reached depending on which anchor set was used in making the comparison. When the common CR item means were examined, Year 2 students appear to be slightly *higher* performing (mean=5.54) than the Year 1 students (5.23.) However, after the rescore, Year 2 students seem to be *lower* performing than the Year 1 students (mean=6.51.) This finding is consistent with the across year rater variation reported earlier. Lastly, based on the common SR items, Year 2 students (mean=7.58) appear to be *lower* performing than the Year 1 students (mean=7.68.) Thus the Year 2 students' performance on the rescored common CR items is more in line with those based on the SR items. Based on these findings, Year 2 students' scores might be inflated if the common CR items were included in the anchor and that rater leniency was left uncorrected. Furthermore, using mixed-item anchors, mixed-item anchors adjusted for rater variation, or SR item anchors are likely to produce different equating results.

Table 4. Summary Statistics for Equating Year 2 and Year 1 Form for Grade 4

Score	Year	Mean	SD	Skewness	Kurtosis
Total Score	Year 2	55.03	19.35	-0.47	-0.77
	Year 1	56.55	18.38	-0.45	-0.73
	Year 1 Rescore	56.42	18.52	-0.42	-0.79
SR+CR anchor	Year 2	9.78	4.59	-0.27	-0.99
	Year 1	9.95	4.41	-0.28	-0.82
	Year 1 Rescore	9.98	4.49	-0.25	-0.9
CR anchor	Year 2	3.8	2.56	-0.03	-1.22
	Year 1	3.77	2.42	-0.06	-1.13
	Year 1 Rescore	3.8	2.52	-0.02	-1.21
SE anchor	Year 2	5.97	2.51	-0.39	-0.67
	Year 1	6.18	2.48	-0.46	-0.64
	Year 1 Rescore	6.18	2.48	-0.46	-0.64

Table 5. Summary Statistics for Equating Year 2 and Year 1 Form for Grade 6

Score	Year	Mean	SD	Skewness	Kurtosis
Total Score	Year 2	72.39	20.95	-0.92	-0.26
	Year 1	72.21	20.34	-0.66	-0.23
	Year 1 Rescore	72.31	20.29	-0.67	-0.23
SR+CR anchor	Year 2	15.38	5.03	-0.78	0.1
	Year 1	15.49	5	-0.62	-0.17
	Year 1 Rescore	15.52	4.89	-0.66	-0.07
CR anchor	Year 2	5.09	2.46	-0.33	-0.59
	Year 1	5.08	2.52	-0.18	-0.73
	Year 1 Rescore	5.1	2.4	-17	-0.63
SE anchor	Year 2	10.29	3.02	-0.98	0.55
	Year 1	10.41	2.91	-0.91	0.43
	Year 1 Rescore	10.41	2.91	-0.91	0.43

Table 6. Summary Statistics for Equating Year 2 and Year 1 Form for Grade 8

Score	Year	Mean	SD	Skewness	Kurtosis
Total Score	Year 2	77.46	21.54	-0.83	-0.12
	Year 1	77.4	21.6	-0.79	-0.24
	Year 1 Rescore	78.62	21.52	-0.83	-0.16
SR+CR anchor	Year 2	13.12	4.66	-0.6	-0.37
	Year 1	12.91	4.56	-0.43	-0.56
	Year 1 Rescore	14.2	4.46	-0.62	-0.38
CR anchor	Year 2	5.54	2.4	-0.4	-0.46
	Year 1	5.23	2.43	-0.13	-0.59
	Year 1 Rescore	6.51	2.36	-0.6	-0.19
SR anchor	Year 2	7.58	2.73	-0.66	-0.47
	Year 1	7.68	2.6	-0.64	-0.41
	Year 1 Rescore	7.68	2.6	-0.64	-0.41

To examine the effect of integrating rater adjustment process in the Stocky-Lord equating procedure, the degree of alignment between the item difficulty of the anchor item parameters and their actual p-values was compared between the two mixed item type linking methods. There are various different ways to evaluate the Stocky-Lord equating procedure. In this study, we chose a whole item approach in order to summarize the overall impact of applying rater adjustment to the equating results. Two items could have somewhat different discrimination, difficulty, and guessing parameters yet still on the whole perform in the same way. Consequently, it makes sense to look at them as a whole rather than to look at its individual item parameters. This approach reflects a whole item approach in evaluating the results of the Stocking-Lord equating procedure. The estimated p-values were computed using the anchor item parameters along with the ability distribution based on the anchor set. The actual p-values were the classical p-

values. The p-values plots when both SR and CR were used as anchors are presented in Figure 1a to Figure 3b. The actual p-values are displayed on the horizontal axis and the estimated p-values are displayed on the vertical axis. A comparison of Figure 1a and Figure 1b showed that the alignment between the actual and the estimated p-values got slightly worse after the rater adjustment. The correlation between the actual and the estimated p-values decreased from 0.95 to 0.92. Such decrease in alignment is evident by examining item 78, which is a CR item. Its estimated p-value was closer aligned with the actual p-value before the rater adjustment. This decrease in alignment between the actual and the estimated p-values for item 78 reflects effect of the rater adjustment in the overall equating results. Figure 2a and 2b showed that for grade 6 there was great alignment between the anchor and the equated parameters and there was no improvement in alignment with the rater adjustment. The correlation between the actual and the estimated p-values was unchanged.

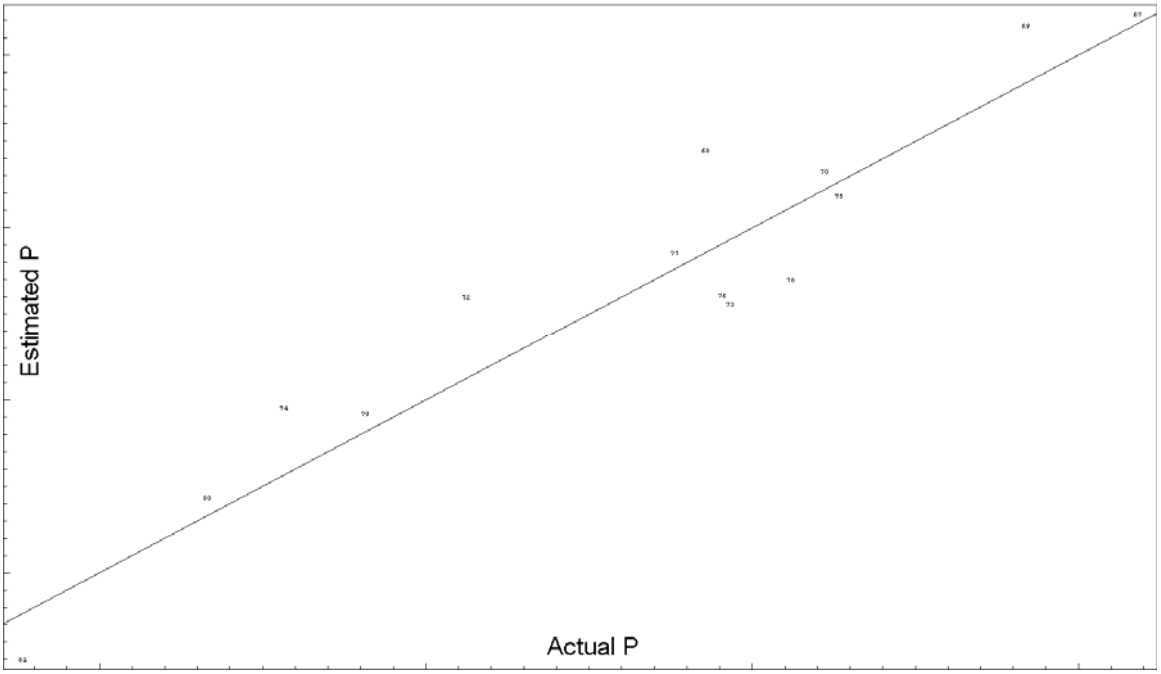
Figure 3a and Figure 3b showed that the overall alignment between the actual and the estimated p-values was relatively poor, with or without rater adjustment, as compared to grade 4 and 6. When the anchor and the equated parameters were produced in two consecutive years, the correlation between two sets of parameters is typically in the 90s. A closer examination of the plots reveals that the degree of misalignment is most severe for two of the CR items, item 91 and item 99. There was a significant improvement in how the actual and the estimated p-values aligned when the rater adjustment was applied. The correlation between the actual and the estimated p-values increased from 0.80 to 0.86. The alignment for item 91 has improved, however, the alignment for item 99 deteriorated after the rater adjustment. Furthermore, the estimated p-values after the rater

adjustment had been applied tend to under-estimate their actual p-values. This is evident by the disproportional number of estimated p-values for both SR and CR items fall below the diagonal line in Figure 3b.

Figure 4 to Figure 6 present the actual and the estimated p-value plot for the three grades when the SR items were used to equate the forms. Since the CR items were not included in the anchor, the alignment between the actual and the estimated p-values improved, especially for grade 8. The correlation between the actual and the estimated p-values are in the high 90s for grade 6 and 8.

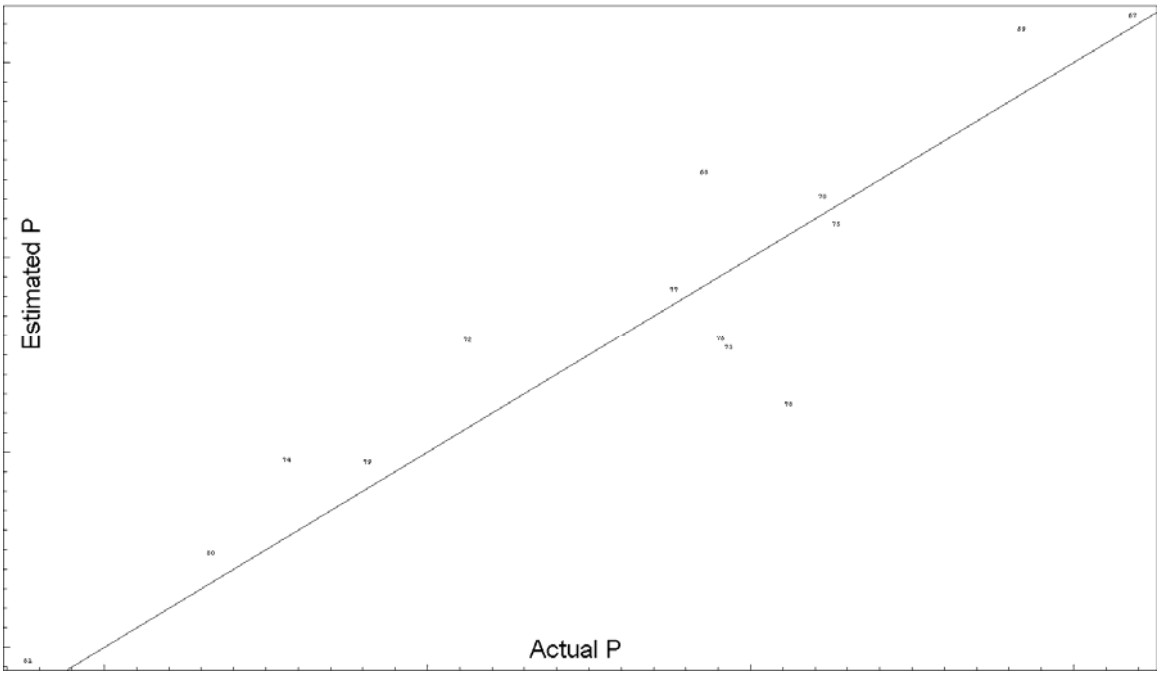
Comparing Figure 6 and Figure 3a, it is evident that the SR anchor items behavior similarly in the two plots. However, when the rater adjustment had been applied through the equating process (Figure 3b), the performance of the SR anchor items changed. Again, it appears that the rater adjustment that had been introduced in the equating process had a significant impact on the stability of the SR anchor items.

Figure 1a. Grade 4 P-value Plots without Rater Adjustment



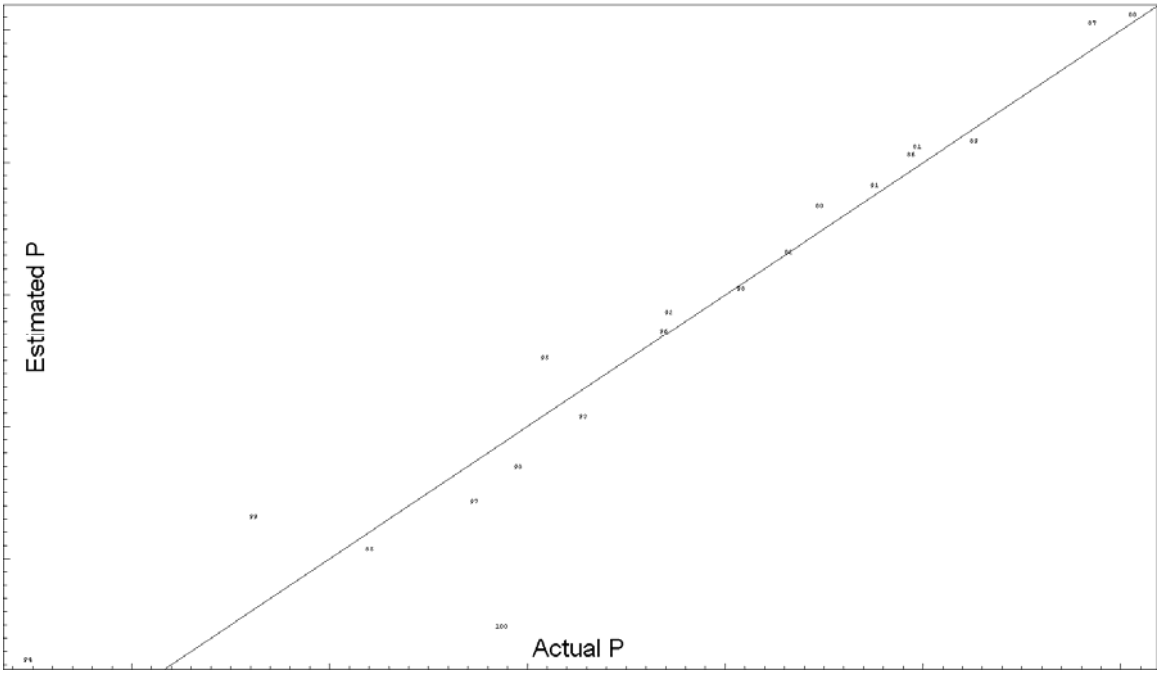
r=0.95

Figure 1b. Grade 4 P-value Plots with Rater Adjustment



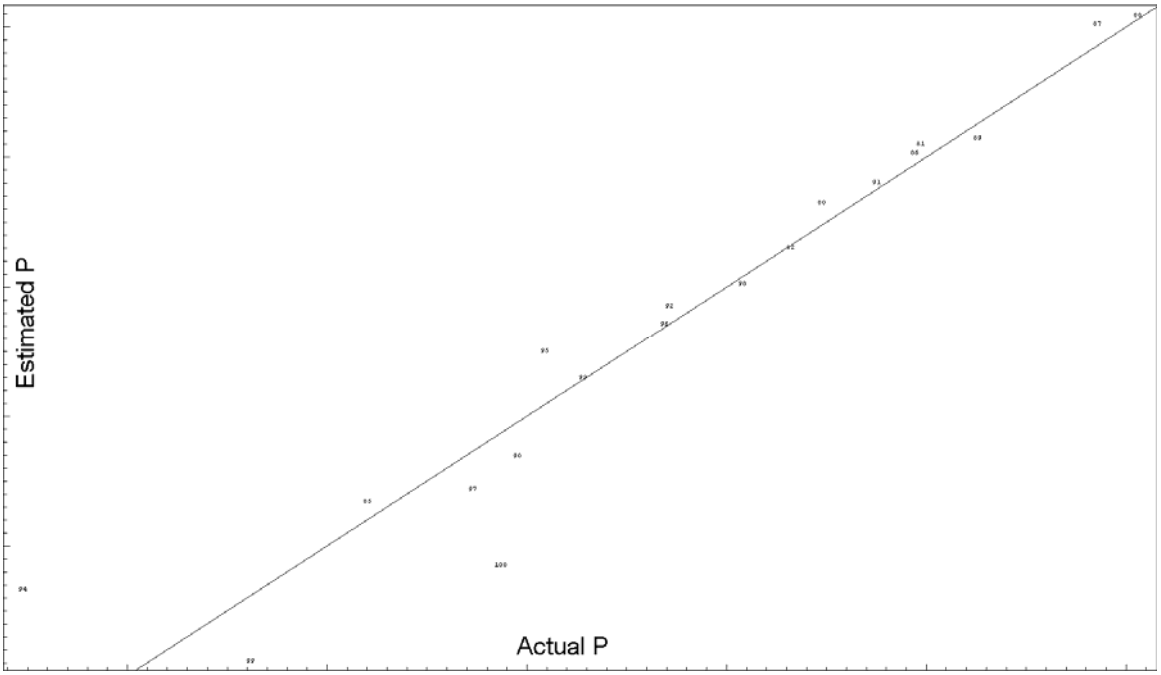
r=0.92

Figure 2a. Grade 6 P-value Plots without Rater Adjustment



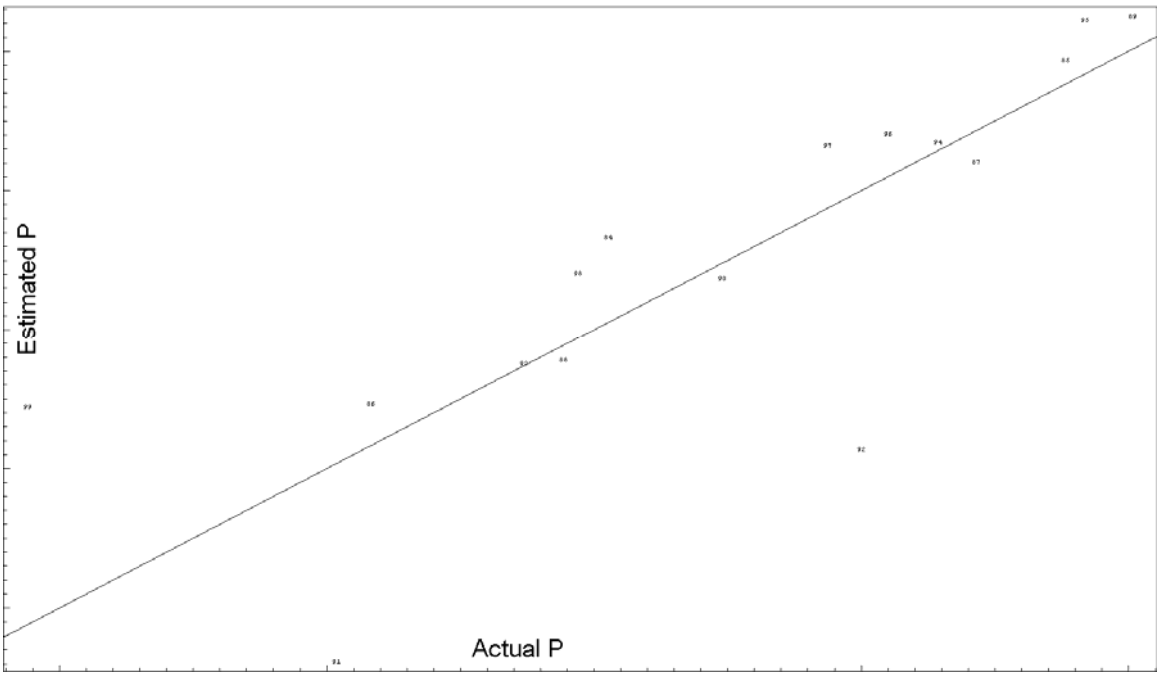
r=0.95

Figure 2b. Grade 6 P-value Plots with Rater Adjustment



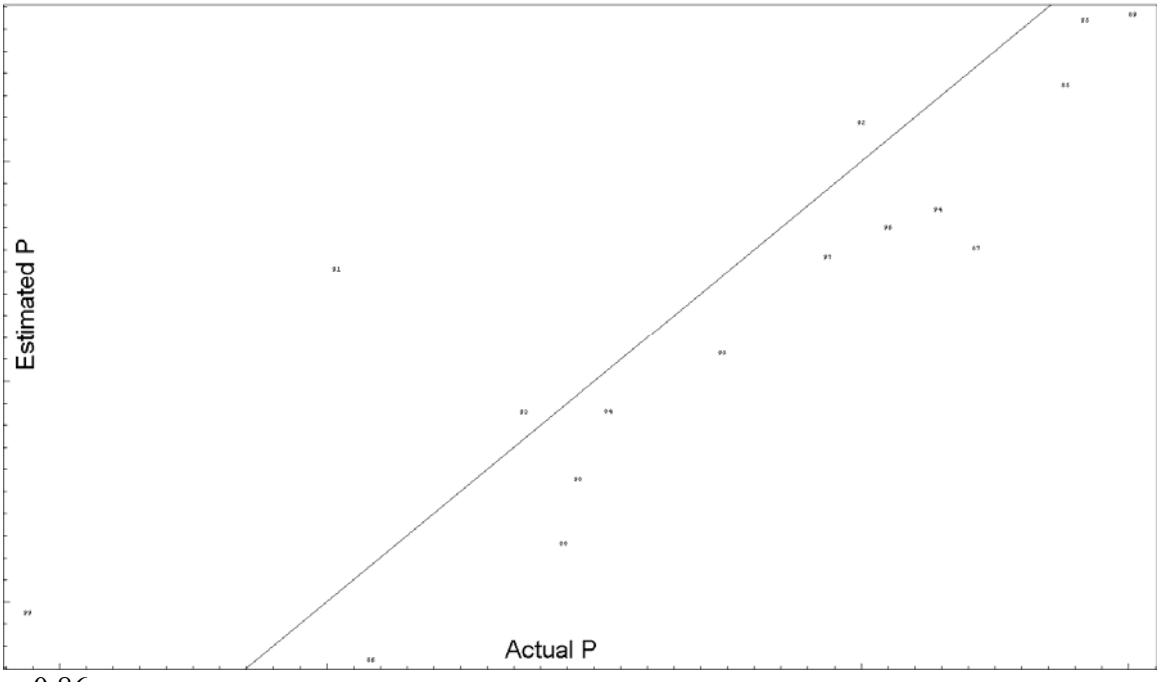
r=0.95

Figure 3a. Grade 8 P-value Plots without Rater Adjustment



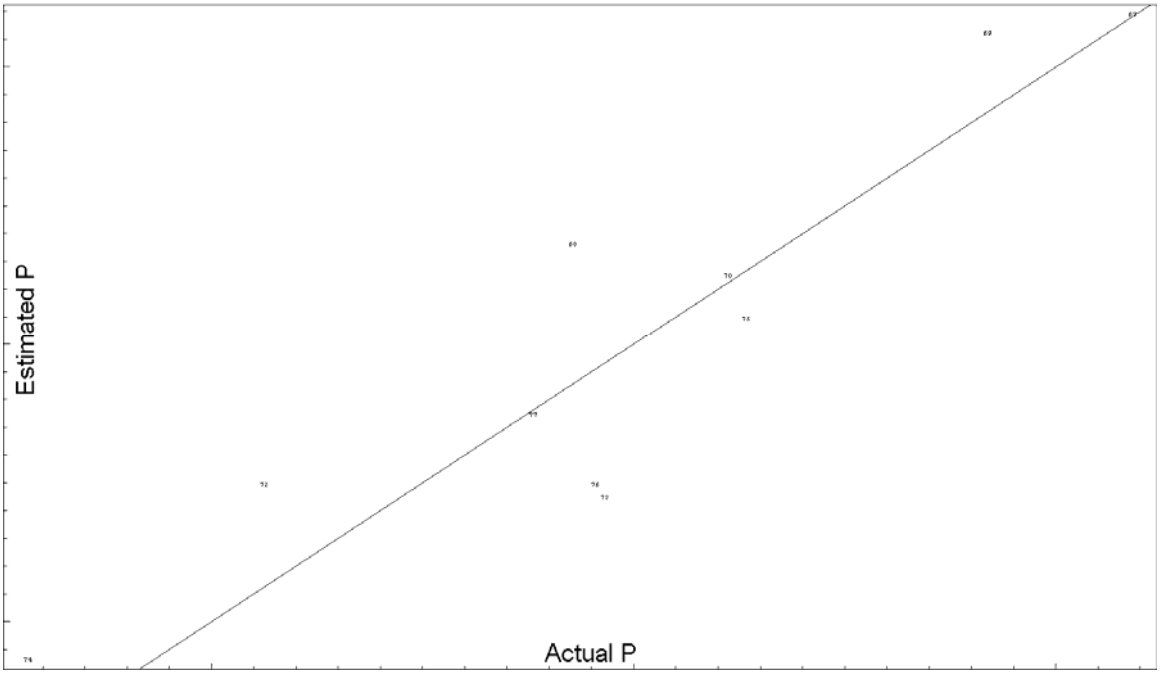
r=0.80

Figure 3b. Grade 8 P-value Plots with Rater Adjustment



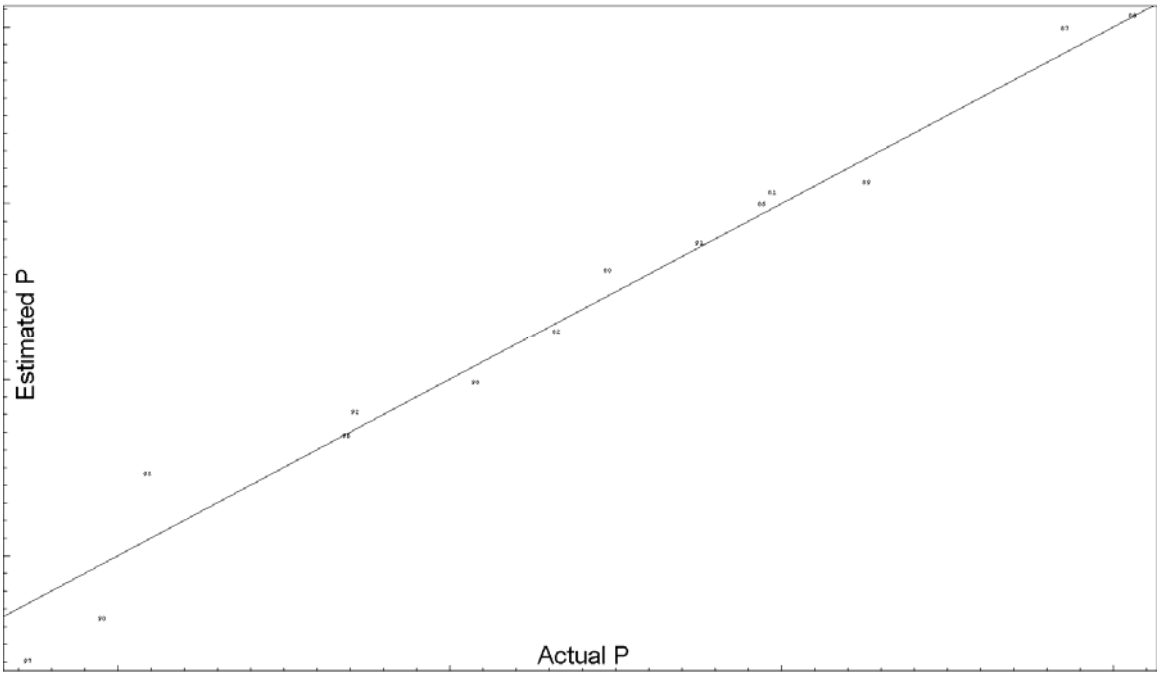
r=0.86

Figure 4. Grade 4 P-value Plots Using SR as Anchor



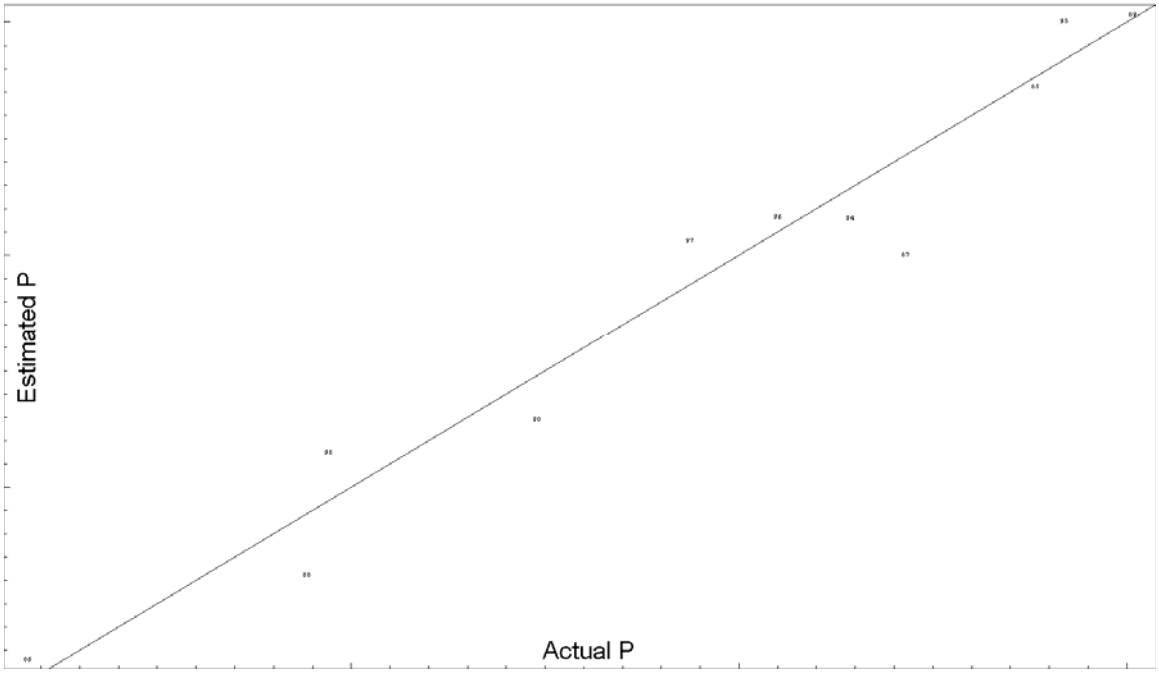
r=.90

Figure 5. Grade 6 P-value Plots Using SR as Anchor



r=0.99

Figure 6. Grade 8 P-value Plots Using SR as Anchor



r=.98

Scoring of the Examinees

The transformed item parameter estimates were used to convert the examinees' scored responses to obtain the number-correct to scale score conversion. The ability estimates were placed on the scale score with a multiplier of 20 and an additive of 200. The highest and the lowest obtainable scale scores were set to 100 and 300. To help visualize the results, the test characteristics curves (TCC) based on the three anchor sets are presented in Figures 7 through 9. In these figures, the number correct scores were rescaled to percent correct to facilitate interpreting the results. The blue line showed the test characteristic curve for Year 2 using mixed-item anchors without rater adjustment, the pink line showed the characteristic curve for Year 2 using mixed-item anchors with rater adjustment, and the green line showed the test characteristic curve for Year 2 using SR items as anchors. The corresponding standard error of measurement curve was displayed at the bottom of each plot. For grade 4 and 6, three TCCs are almost indistinguishable. The standard errors of measurement curves produced by the three procedures are also similar. For grade 8, however, the differences in the TCCs produced by the three different procedures were more pronounced especially between the TCCs with and without rater adjustment. Although not shown on the plots, it was found that eighth graders who received 50 percent of the maximum possible score on the test will be assigned a scale score that is about eight points lower if the rater effect had been adjusted. Furthermore, it appears that using SR anchors produced more accurate results as compared to those using mixed-item anchors without rater adjustment. Turning to the standard error curves produced by the three procedures; it seems that adjusting rater

variation had a negative impact on the precision of estimation for the higher performing students but had a positive impact on the precision of estimation for the lower performing students.

Figure 7. Test Characteristic and Standard Error Curves for Grade 4

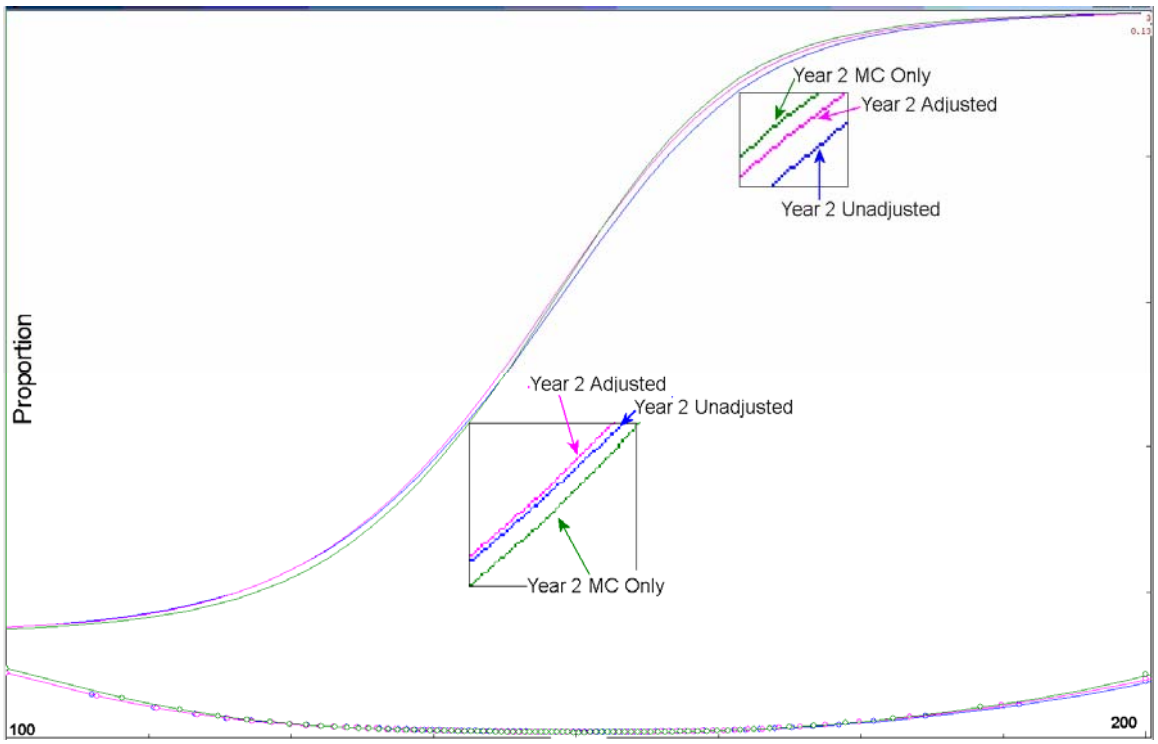


Figure 8. Test Characteristic and Standard Error Curves for Grade 6

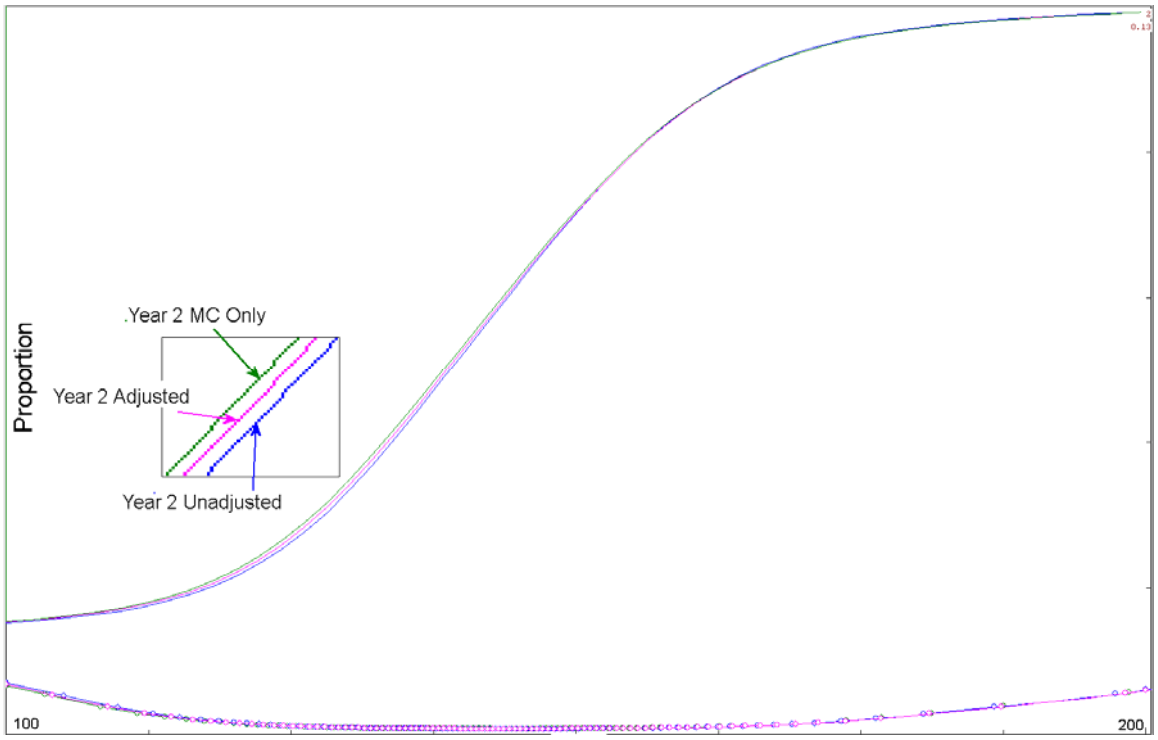
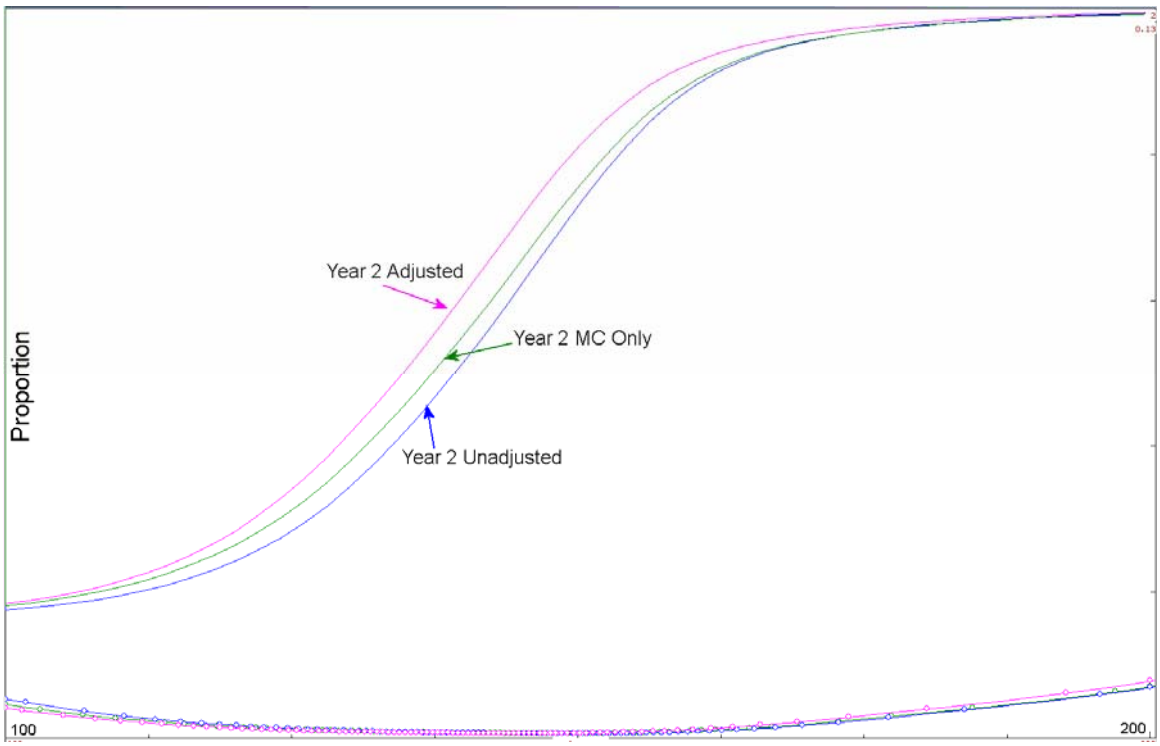


Figure 9. Test Characteristic and Standard Error Curves for Grade 8



Looking more closely at the effect of applying rater adjustment in the equating to the students' scale score distributions, Tables 7 presents the percentage of Year 2 students who were classified as proficient for the three procedures as well as the percent of Year 1 students being classified as proficient. It should be noted that both number-correct scale score and the item-pattern scale score were derived for the examinees. Since the differences between the two scale score distributions were very small across grades and comparisons and that item-pattern scoring is a more commonly used procedure than number-correct scoring, the former was used in this analysis. A somewhat arbitrary cut score of 200 was set to classify students into proficient or not proficient category. For grade 4, using the mixed-item type anchors with rater adjustment in the equating resulted in 2% less students being classified as proficient as compared to those without rater

adjustment. The difference of 2% is relatively small and might be considered random errors. Note that the result based on the SR anchors is exactly the same as those based on mixed-item anchors with rater adjustment; therefore, it could be argued that the three procedures produced the same results.

There was virtually no difference in the percent of students being classified as proficient for grade 6 among the three procedures. Such findings are not surprising given the similarity in the TCCs across procedures.

For grade 8, there was quite a significant decrease (14%) in the percent of students classified as proficient using mixed-item anchors adjusted for rater variation and a smaller but still sizable decrease (4%) in the percent of students classified as proficient using SR anchors. It seems that not accounting for rater variation can produce misleading results under the presence of significant rater variations. Table 8 to Table 10 presents the summary statistics of the students' scale score distributions for each grade by year and by procedure. Overall, the mean and the standard deviation are quite similar across the three procedures with the exception of grade 8. For grade 8, it seems that small differences in means among the three procedures might correspond to a sizable difference in the percent of students being classified as proficient.

Table 7. Scale Score Distributions by Grade

	Percent Proficient (SS≥200)			
	Year 1	Year 2 Mixed-Item Anchor	Year 2 Mixed-Item Anchor With Rater Adjustment	Year 2 SR Only Anchors
Grade 4	56.5%	54.0%	52.1%	52.1%
Grade 6	52.0%	51.5%	51.5%	51.5%
Grade 8	55.5%	58.3%	44.2%	54.2%

Table 8 Grade 4 Reading Scale Score Summary Statistics

	Year 1	Year 2 Mixed-Item Anchors	Year 2 Mixed-Item Anchors With Rater Adjustment	Year 2 SR Only Anchors
Sample Size	1019	2129	2129	2129
Mean	201.11	199.34	198.10	198.34
Std. Error of the Mean	0.800	0.614	0.600	0.575
Median	202.9	202.7	201.4	201.5
Std. Deviation	25.55	28.32	27.70	26.52
Confidence Interval (95%)				
Lower Bound	199.54	198.14	196.92	197.22
Upper Bound	202.69	200.54	199.28	199.47

Table 9 **Grade 6 Reading Scale Score Summary Statistics**

	Year 1	Year 2 Mixed-Item Anchors	Year 2 Mixed-Item Anchors With Rater Adjustment	Year 2 SR Only Anchors
Sample Size	1015	2061	2061	2061
Mean	198.36	197.40	197.21	196.81
Std. Error of the Mean	0.815	0.562	0.571	0.575
Median	201.0	200.5	200.4	200.0
Std. Deviation	25.97	25.52	25.92	26.10
Confidence Interval (95%)				
Lower Bound	196.77	196.29	196.09	195.68
Upper Bound	199.96	198.50	198.33	197.94

Table 10 **Grade 8 Reading Scale Score Summary Statistics**

	Year 1	Year 2 Mixed-Item Anchors	Year 2 Mixed-Item Anchors With Rater Adjustment	Year 2 SR Only Anchors
Sample Size	1001	2156	2156	2156
Mean	202.16	202.79	194.45	200.36
Std. Error of the Mean	0.876	0.587	0.598	0.618
Median	203.3	204.6	196.3	202.2
Std. Deviation	27.72	27.27	27.78	28.68
Confidence Interval (95%)				
Lower Bound	200.44	201.64	193.28	199.15
Upper Bound	203.88	203.94	195.63	201.57

Conclusion

This study illustrated that the rater variation across year could be significant and that such variation warrants the need for making statistical adjustment. The polytomous IRT-based linking method proposed by Tate (1999, 2000) provides a way to adjust rater effect in the equating process under the framework of non-equivalent group common-item equating design. This method was compared with two alternative methods: traditional IRT linking study that links the test forms using a) all the common items without the rater adjustment and b) the common selected response items. The differences in alignment between the anchor and actual p-values were compared among the three methods. The differences in the test characteristic curves and the students' scale score distributions were also compared.

Significant shifts in the parameters after rater adjustment were found for one (grade 8) of the three grades examined. The p-values and TCCs shifted across years when adjusted for rater effects. The impact of the parameter shifts and TCCs manifested in the changes in the proficiency classification before and after adjustment.

The results of this study have important implications to the valid interpretation of the scale-based results. Though time is a constraint in operational testing programs, rater studies need to be more routinely done to examine the direction and the magnitude of the rater variation. Ideally, CR items can be incorporated into anchor sets resulting in better year-to-year equating in mixed item format tests as long as the rater variation can be adjusted through the equating process. Otherwise, year-to-year growth and rater drift might be inextricably confounded. In the case when it is not feasible to apply rater

adjustment in the equating process, using the common SR items is a better approach than using the combined SR and CR items without adjusting for rater effect.

Further studies need to be conducted for year-to-year test administrations that go beyond two administrations and for multi-categorical proficiency classifications with more than two levels. Furthermore, research will need to be conducted to investigate these effects on multiple cut scores and various levels of proximities to these cut scores.

References

Burke, G. R. (2002). PARDUX. Version 6.1 [Computer Software]. Monterey, CA: CTB/McGraw-Hill.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*, 171-191.

Fitzpatrick, A. R., Ercikan, K., Yen, W., Ferrara, S. (1998). The consistency between raters scoring in different test years. Applied Measurement in Education, 11, 195-208.

Ito, K. Sykes, R. (1998). Effects of rater severity and leniency on test forms containing mixed item types. Paper Presented at Annual Meeting of the National Council on Measurement in Education, San Diego, 1998.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331-245.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

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

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press. (Originally published 1960)

Patz, R. J. & Junker, B. W., & Johnson, M. S. (1999). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

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

Tate. R. L. (1999). A cautionary note on IRT based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336-346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, *37*, 329-346.

Wilson, M. & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement*, *19*, 51-71.