The Effect of Year-to-Year Rater Variation on IRT Linking

Shu Jing Yen
Charles Ochieng
Hillary Michaels
Greg Friedman

CTB/McGraw-Hill

Paper Presented at the Annual Meeting of the American Educational Research Association, Montréal, Canada, 2005.

Abstract

Year-to-year rater variation may result in constructed response (CR) parameter changes, making CR items inappropriate to use in anchor sets for linking or equating. This study demonstrates how rater severity affected the writing and reading scores. Rater adjustments were made to statewide results using an item response theory (IRT) methodology based on the work of Tate (1999, 2000). The common item equating design was used to place the second year scores to the first year scores after a re-score of the first year test in order to adjust for rater effects. Two samples of data from contiguous years, designated as Year 1 (n ~ 1,200) and Year 2 (n ~ 2,000), from the writing and reading portions of a statewide assessment were examined. The writing test consisted of 32, 36, and 40 selected-response items for grade 4, 6, and 8 and a single writing prompt scored on a six-point scale (0-5) scored by two raters whose scores are added for a composite. The reading test consists of 75, 93, and 91 selected-response items and 12, 14, and 16 constructed response items for grade 4, 6, and 8, respectively. All the CR items in reading were scored on a three-point scale (0-2.)

The resulting item parameters were compared between year one and two, with and without rater adjustment. For writing, there were significant shifts in the parameters after rater adjustment. 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 the study suggests that raters were not consistently more severe or more lenient between grades or content areas, 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 students scores (Engelhard, 1992, Lunz, Wright, & 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 examine the effect of the year-to-year variation on IRT linking. Traditional IRT linking studies were used to link test forms using common items then a modified 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 the same between the two years.

Method

Data from the writing and reading portion of a statewide 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 writing test consisted of 32, 36, and 40 SR items for grade 4, 6, and 8 and a single writing prompt for all three grades. The writing prompt was scored on a six-point scale (0-5). Two independent raters scored the student's response. The student's final writing score was the composite of the two independent ratings with the maximum possible score of 10. The reading 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 in reading 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. Note that by the nature of the test design, only a single extended writing prompt with 11 levels (0-10) was used as the anchor while for reading, the anchor item block consists of both the SR and the CR items.

Table 1. The Number of Score Points by Item Type

	Non-Anchor		Anchor	
	SR	CR	SR	CR
Writing				
Grade 4	32(76%)	0(0%)	0(0%)	10(24%)
Grade 6	36(78%)	0(0%)	0(0%)	10(22%)
Grade 8	40(80%)	0(0%)	0(0%)	10(20%)
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 is 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 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 are 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 toward the end of the scoring session to rescore the Year 1 student papers no 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 year tend to be small for reading. The exceptions are item 99 in grade 6 and items 91, 92, and 93 in grade 8. The generally small differences in means suggest that the 1st- and 2nd-year raters, on average, gave very similar scores. However, on average, the year 2 raters tended to be more lenient than the year 1 raters for the reading CR items. In writing, the year 2 raters tended to be more lenient in grades 4 and 8 but more severe in grade 6. The differences in raw score mean and standard deviations between the 1st and 2nd years tended to be large for grade 4 and 6 and small for grade 8. For grade 6, the difference in raw score mean between the 1st and 2nd year raters is especially large (0.43.)

The percentage of perfect, adjacent and discrepenent agreement between years and weighted kappa is presented in Table 3. Only the rater agreement statistics for the first read of each year was used in reporting since the second year statistics were very similar. For writing, perfect agreement rates varied from 44% to 48%. Consistent with the grade 6 writing raw score statistics, item 37 has the lowest perfect agreement as compared to items in grade 4 and grade 8. For reading, the percent perfect agreement varied from 40% (grade 8, item 91) to 89% (grade 4, item 79). Grade 8 reading tended to have lower perfect agreement rates compared to grades 4 and 6.

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
Writing	Grade 4	Item 33	6.66	6.85	-0.19	1.96	1.66
	Grade 6	Item 37	6.90	6.47	0.43	1.68	1.69
	Grade 8	Item 41	6.84	6.91	-0.07	1.56	1.70

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
Writing	Grade 4	Item 33	46	42	11	0.87	0.85
	Grade 6	Item 37	44	43	12	0.87	0.83
	Grade 8	Item 41	48	41	11	0.84	0.84

Analysis

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.) All of the items within a grade and content area combination were simultaneously scaled using PARDUX software program (Burket, 2002.)

To provide a baseline for evaluating the proposed linking design that adjusts for year-to-year rater variation, the traditional linking method was first applied. Each Year 2 form was equated to the Year 1 form scale using a set of anchor items. The Stocking-Lord equating method (Stocking & Lord, 1983) was used to place the Year 2 test on the Year 1 scale using all the common items between the forms. To adjust for year-to-year rater variation in the equating, a two-step procedure was used. First, the Year 1 item parameters for the CR items were re-estimated using the ratings assigned by the Year 2 raters. Through this recalibration process, the Year 1 item parameters for the CR items were adjusted for rater variation. Each Year 2 form was again equated to the Year 1 form scale using the anchor items. Second, the Stocking-Lord 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 Step 1, the changes in the constructed response anchor item parameters are reflective of changes in item difficulty as opposed to a combination of rater variation and item difficulty.

Results and Discussions

To evaluate the difference between the equating results produced by the two designs, one without rater adjustment and the other with rater adjustment, item parameters were first compared. Because there was only one anchor item used to equate the Year 1 and Year 2 writing, it is fairly easy to compare the parameter estimates produced by the two designs. Table 4 reports the final item parameter estimates of the anchor item after equating with and without rater adjustment. For grades 4 and 6, the differences in item discrimination parameters between the two designs are negligible. For grade 8, the item discrimination parameter became larger after the rater variation has been adjusted. For grade 4 and 8, the b parameters decreased after the rater adjustment. The implication is that without the rater adjustment, the b parameters would be over estimated. For grade 6, however, the b parameters increased after the rater variations had been adjusted indicating that the difficulty parameter would be underestimated without the rater adjustment.

Since the number of anchor items used to equate the Year 1 and Year 2 reading assessment are relatively large as compared to those for writing assessment, the reading item parameter estimates produced by the two designs were not reported in this paper. However, detailed results related to the Stocking-Lord equating for reading assessment are available in Yen et al (Yen, O'chieng, Michaels, & Friedman, 2005.)

Table 4

Comparison of item parameters with and without adjustment for Writing by grade

Grade	Grade 4		Grade 6		Grade 8	
Parameters	With Adjustment	Without Adjustment	With Adjustment	Without Adjustment	With Adjustment	Without Adjustment
Item	33	33	37	37	41	41
Levels	11	11	11	11	11	11
a	0.340	0.360	0.490	0.480	0.420	0.520
b_1	-0.280	-0.410	-1.410	-1.250	-0.850	-0.910
b_2	-1.920	-2.050	-1.900	-1.750	-1.910	-1.960
b_3	-0.460	-0.590	-1.120	-0.960	-0.680	-0.740
b_4	-1.510	-1.650	-1.780	-1.620	-1.520	-1.580
b_5	-0.010	-0.140	-0.130	0.030	-0.510	-0.570
b_6	-0.360	-0.500	-0.880	-0.720	-0.760	-0.820
b_7	0.600	0.470	1.000	1.160	0.970	0.910
b_9	0.640	0.510	0.420	0.580	0.710	0.650
b_9	1.110	0.980	1.190	1.350	1.230	1.180
b_{10}	0.860	0.730	1.240	1.400	1.550	1.490
Mean of b-parameter	-0.133	-0.265	-0.337	-0.178	-0.177	-0.235

Scoring of the Examinees

The final 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) for the two designs are presented in Figures 1 to 6. In these figures, the number correct scores were rescaled to percent correct to facilitate interpreting the results. For writing, the differences in the TCCs produced by the two designs are quite dramatic for grade 4 and 6. For example, six graders who received 50 percent of the maximum possible score on the test will be assigned a writing scale score that is about six points higher if the rater effect has been adjusted. For reading grade 4 and 6, the two TCCs are almost indistinguishable. For grade 8 reading, however, the differences in TCCs produced by the two designs are more pronounced. Eighth graders who received 50 percent of the maximum possible score on the test will be assigned a reading scale score that is about eight points lower than if the rater effect has been adjusted

Figure 1. Grade 4 Writing TCCs

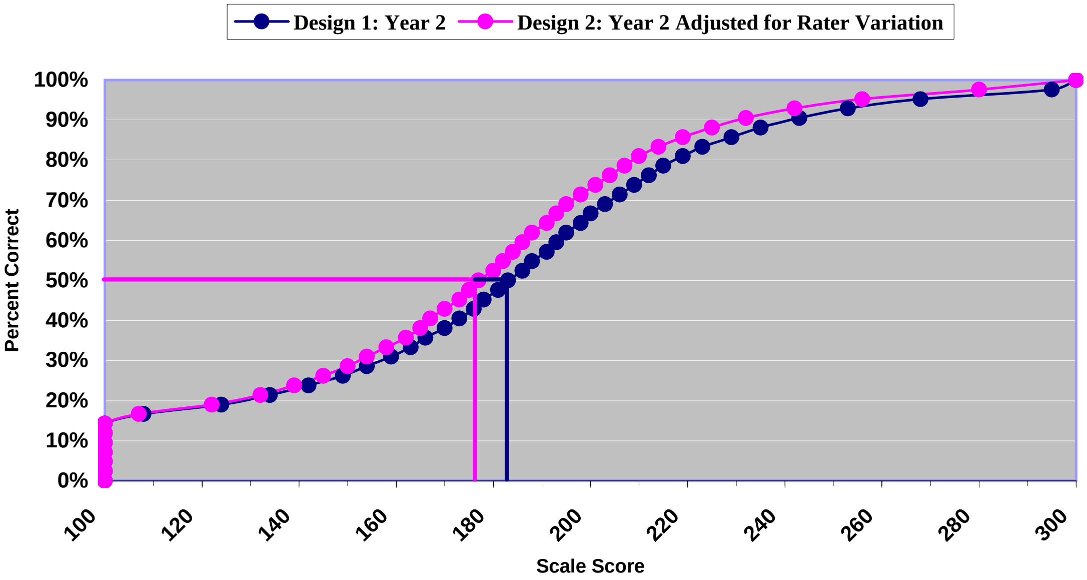


Figure 2. Grade 6 Writing TCCs

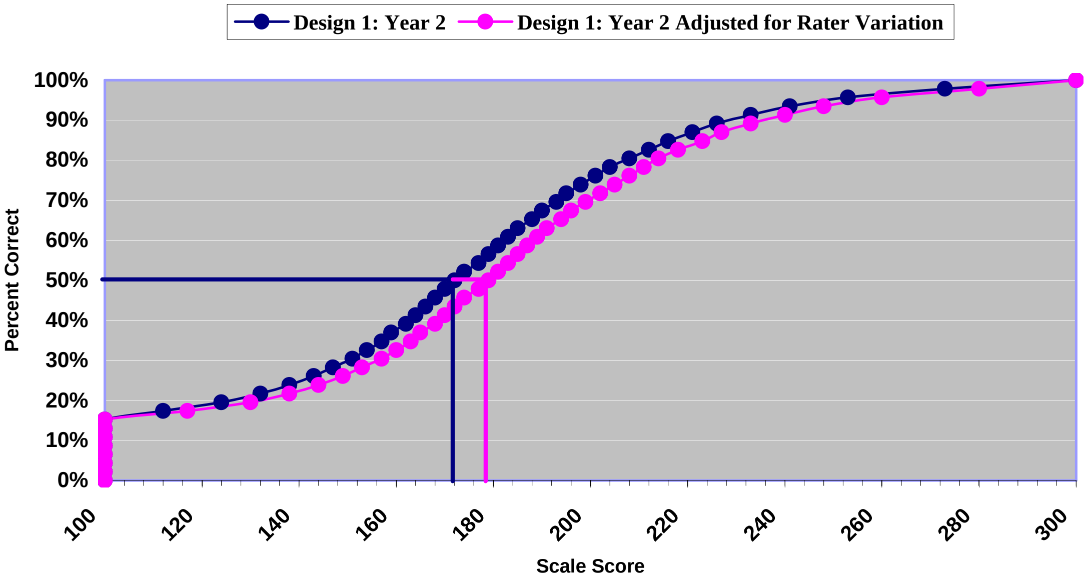


Figure 3. Grade 8 Writing TCCs

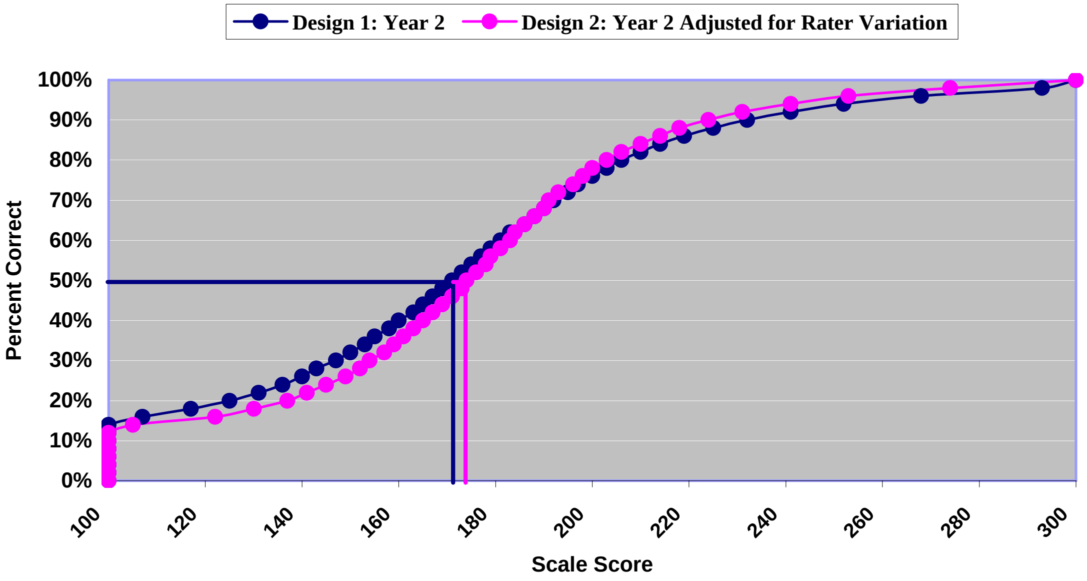


Figure 4. Grade 4 Reading TCCs

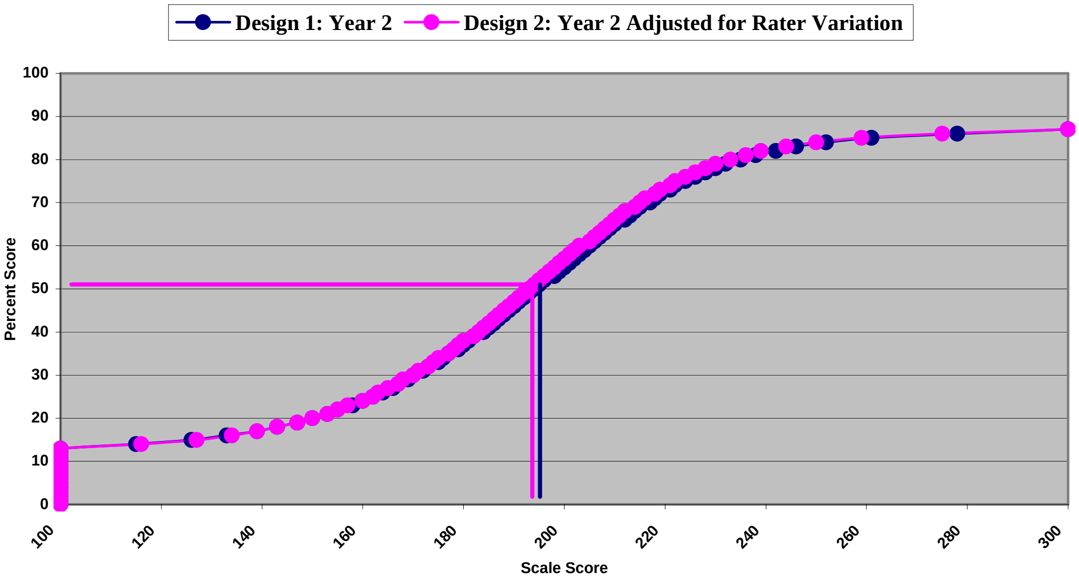


Figure 5. Grade 6 Reading TCCs

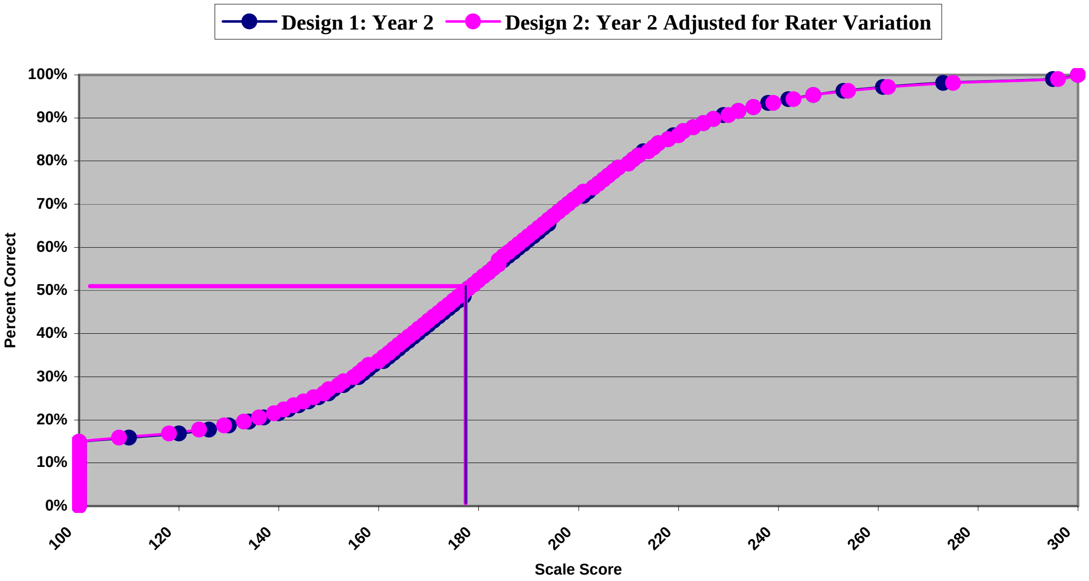
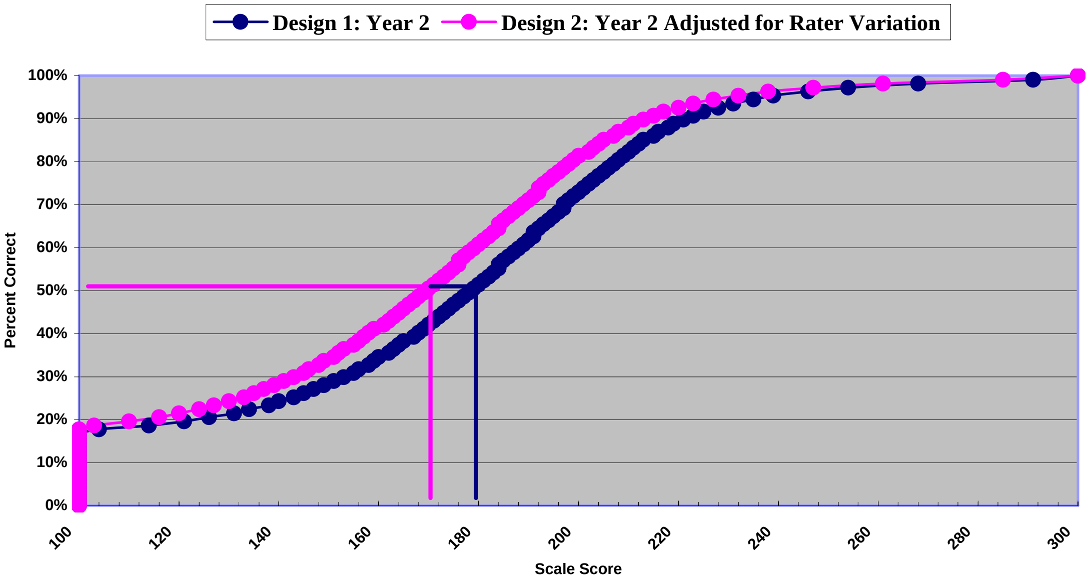


Figure 6. Grade 8 Reading TCCs



The scale score distributions produced by the two designs were examined further. Tables 5 and 6 present percentages of students who were classified as proficient before and after the rater adjustment. A somewhat arbitrary cut score of 200 was set to classify students into proficient or not proficient category. Table 5 shows that there was a significant impact of the percent of students being classified as proficient after the rater variation has been adjusted across all three grades for writing. Grade 4 shows the most dramatic impact (12%) as compared to grade 6 and 8.

Table 6 shows that for reading, there was no change in the percent of students being classified as proficient for grade 6, a small change for grade 4, and a quite significant change (14%) for grade 8.

Table 5
Writing Scale Score Distributions Before and After Rater Adjustments

	Percent Proficient (SS≥200)	
	Before Rater Adjustment	After Rater Adjustment
Grade 4	54.4%	42.1%
Grade 6	41.6%	52.0%
Grade 8	47.8%	43.0%

Table 6
Reading Scale Score Distributions Before and After Rater Adjustments

	Percent Proficient (SS≥200)	
	Before Rater Adjustment	After Rater Adjustment
Grade 4	54.0%	52.1%
Grade 6	51.5%	51.5%
Grade 8	58.3%	44.2%

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 IRT 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 study showed how rater severity affected the writing and reading scores before and after rater adjustments were made. By using the scoring tables and examining the students above and below an arbitrary cut score, demonstrates how not accounting for rater severity can produce misleading results. This has 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 so that CR items can be incorporated into anchor sets resulting in better year-to-year equating in mixed item format tests.

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.

Yen, S.J, O'chieng, C., Michaels, H, & Friedman, G (2005) Adjusting for year to year rater variation in IRT linking. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Canada, 2005.