A Comparison of Pre-Equating and Post-Equating using Large-Scale Assessment Data

Ye Tong
Sz-Shyan Wu
Ming Xu

Paper to be presented at the American Educational and Research Association annual conference in New York City, March 2008

Introduction

Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004), even though the test forms consist of different items. Various equating models and procedures have been developed in the literature. Equating is widely used in various large-scale assessments where people want to make meaningful comparisons of test scores across administrations.

There are many types of equating methodologies. In terms of design, there are randomly equivalent groups design and common-item non-equivalent groups design (Kolen & Brennan, 2004). In terms of testing models, there are classical test theory based equating model and item response theory (IRT) (Lord, 1980) based equating model. In terms of where equating fits into the assessment cycle, there are pre-equating and post-equating. The contrast between pre-equating and post-equating is the focus of the current paper.

Pre-equating, as the term suggests, is to conduct equating prior to the operational testing. Post-equating, similarly, is to conduct equating after the operational testing. Both equating processes are widely used in K-12 large-scale assessment programs. To conduct pre-equating, statistical procedures are applied typically to field test data. With post-equating, statistical adjustment of test difficulty is applied to operational test data.

Both of these equating models are used extensively with the state assessment programs. Each of the two models has its own advantages and disadvantages. The use of pre-equating, when applicable, can facilitate the operational process in terms of rapid score reporting, more time for quality control and more flexibility in the assessment. On the other hand, a variety of issues need to be considered when using pre-equating in practice. In fact, many researchers have voiced concerns about using pre-equating. Although item parameters have been found to be acceptably accurate under pre-equating (Bejar and Wingersky, 1982), problems can occur when the field test items are presented in a separate section (Eignor, 1985; Eignor and Stocking, 1986; Stocking and Eignor, 1986; Kolen and Harris, 1990), because of the concerns on students' motivation. The use of post equating, when applicable, does not have the same motivation concern sometimes occurring with pre-equating. Also, post-equating uses post administration data and is sometimes considered to yield more accurate analysis results. On the other hand, when the reporting window is extremely tight, as the case with some graduation tests or end-of-course tests in various states, post-equating has to occur within a very short time window and hence less time for equating process and quality control. Also, post-equating always has to require testing data be collected and therefore, rules out the possibility of classroom teachers rating the exams and reporting the scores.

Much research has been conducted in an attempt to compare pre-equating and post-equating and how they tend to impact the equating results. Pre-equating continues to be used by many large-scale assessment programs due to various practical and policy reasons. Under such circumstances, how can pre-equating be conducted and what are the modifications researchers and psychometricians can do to enhance the accuracy of pre-equating results? In this study, the authors compare pre-equating and post-equating in

various subject areas and where large discrepancy exists, the authors attempt to provide rationale and possible solutions to decrease the differences between different equating models.

Data

The large-scale assessment data from a state program were used in the current study. A pre-equating model using the IRT Rasch (Rasch, 1960) and the Partial Credit Model (PCM, Masters, 1982) model is used, because of the necessity to have scoring tables prior to test administration. When the items are field tested, it is a stand-alone session. Therefore, students' motivation on the field test tends to be less than optimal. In this study, item parameter estimate and the raw score to theta (e.g., scoring table) relationship for pre-equating model were calibrated and developed on the field test data.
In this separate field testing, researchers are concerned that students may not try as hard as they would have if the field test items are embedded within an operational test. Because of the lack of motivation, items may appear more difficult than they actually are and therefore, lead to somewhat less optimal item parameter estimation.

In this study, post-equating was also conducted and the results were compared with the pre-equating results to further observe whether lack of motivation in field testing has impacted the operational scoring. Post-equating was conducted on a representative sample obtained through post administration score collections. A representative sample of the state population, in terms of ethnicity, gender, social economic status and geographic location, as well as assessment performance, was obtained. Both pre and post-equating were carried out in three subject areas: mathematics, science and social studies. These three subjects were used because of their unique features in the field testing and pre-equating as well as the findings with the pre-post equating contrast.

For each of the three subjects, there are a mixture of multiple choice (MC) items and constructed response (CR) items, and sometimes, essay questions. Table 1 reports the number of MC items, number of CR items, number of essay questions and the sample size for each of the three subjects.

As can be observed from Table 1, the sample size is fairly large for each content area. There is a good mix of MC items and CR items. The social studies test also has two essay questions. With the mathematics test, the CR items have maximum score points ranging from 2 to 6 and the MC items are weighted by 2 in the final scoring. With the science test, the CR items have a maximum score point of 2 and no additional weighting is applied. With social studies, CR items have a maximum score point of 2; the essay questions have a maximum score point of 5 and are also weighted by 3 in the final scoring. The sample sizes were fairly large.

Methodology

Pre-equating

In the pre-equating process, "mini" field test forms were constructed, typically with 20-25 items per form. There were also "anchor forms" that were used to equate all field test items onto the operational scale. A representative sample is identified throughout the state and these mini-forms were spiraled within the classroom so that the groups of students taking each form were randomly equivalent. The field test forms were equated using two designs: equivalent groups and common item. Specifically, the equivalent groups design was used for mathematics and science. Both the equivalent groups and the common item equating designs was used with social studies.

In this study, the pre-equating process was accomplished using the following steps:

Step 1: Obtain Rasch and PCM item parameter estimates using the field test data.

Step 2: Rasch and PCM item parameters are placed onto the operational scale through anchor items and above equating designs.

Step 3: Operational test items are selected from the field test item bank (e.g., pre-equated item parameters) based on content coverage and the average Rasch test difficulty.

Step 4: A raw score to theta relationship (e.g., scoring table) for the operational test form is developed using the field test pre-equated item parameters.

Weighting of the MC items or essay items were not applied at the calibration phase, but at the scoring phase. Also, when constructing test forms across years, when computing the average test difficulty, items were also weighted by their maximum score points and scoring weights.

Post-equating

In post-equating, the post operational item parameters and scoring table were produced using the operational data. To conduct a pre-post contrast, the key is to try to replicate pre-equating using operational data. Therefore, during post equating, all the rules used in pre-equating, when applicable, were also followed during post-equating. The mean/mean equating method (Kolen & Brennan, 2004) was applied to place the item parameter estimates and scoring tables on the same scale when comparing the pre and post equating models. With these three subjects, specifically, no weighting was applied during calibration phase, but at the scoring phase. The following steps were followed during post-equating:

Step 1: Calibrate all items on the operational form allowing the post operational item difficulties to center at a mean value of zero and obtain raw score to theta scoring table.

Step 2: Obtain average Rasch test difficulty using the post operational item parameters from the previous step, taking into account both maximum score points and scoring weights.

Step 3: Obtain the scaling constant for post-equating by subtracting the average Rasch item difficulty from step 2 from the average Rasch item difficulty from pre-equating.

Step 4: Adjust all the post operational item parameters by adding the scaling constant obtained from step 3. Also adjust the scoring table (step 1) by adding the scaling constant to place the scoring table on the same scale as the pre-equating scoring table.

Evaluation Criteria

One obvious feature to observe is item parameter estimates. After placing the item parameters from post-equating onto the pre-equating operational scale, Rasch item difficulty values can be contrasted. In most applications of the Rasch model, correlations between item parameter estimates obtained between two administrations are expected to be above 0.90 and average absolute differences between estimates are expected to be below 0.20. The same criteria can be applied when comparing pre-post results.

It is also important to observe how different the raw score-to-theta scoring tables tend to be based on pre-post contrast. In large-scale assessment context, decisions on classifications are also very important. In this study, percentages of students in each of the performance levels are also contrasted between pre and post equating.

Another reliability index we looked at is classification accuracy. Namely, based on the Rasch model, what percentage of students is accurately classified. Rudner (2005) method was applied to compute classification accuracy index for both pre and post equating results. To calculate the classification reliability index under the Rasch model for a given ability score θ. the observed score $\hat{\theta}$ is expected to be normally distributed with a mean of θ and a standard deviation of $SE(\theta)$ (the SEM associated with the given θ). The expected proportion of examinees with true scores in any particular level is

$$\text{PropLevel}_k = \sum_{\theta=cut_{\theta_c}}^{cut_{\theta_d}} \left(\phi\left(\frac{cut_{\theta_b} - \theta}{SE(\theta)} \right) - \phi\left(\frac{cut_{\theta_a} - \theta}{SE(\theta)} \right) \right) \varphi\left(\frac{\theta - \mu}{\sigma} \right),$$

where cut_{θ_a} and cut_{θ_b} are Rasch scale points representing the score boundaries for levels of observed scores, cut_{θ_c} and cut_{θ_d} are the Rasch scale points representing score boundaries for levels of true scores, ϕ is the cumulative distribution function of the

achievement level boundaries, and φ is the normal density function associated with the true scores (Rudner, 2005).

Results

Math

Table 2 presents item parameter estimates contrast between pre and post equating results. The first two columns report p-values from stand-alone field testing and operational testing. As can be observed from this table, generally speaking, the p-values tend to be higher for operational testing, especially for items with more points such as items 33 and 34. The reason for this is possibly because at stand-alone field testing, students tend to be less motivated to try their best to answer test questions. For items with high score points, students tend to be even less willing. In fact, for items with maximum score points of 6, such as items 33 and 34, the response rate from field testing tends to be so low that it is hard to estimate their Rasch item difficulty. In order to produce scoring tables for pre-equating, average item difficulty for these six-point items in the item bank is used instead. As Table 2 indicates, the average value used for the two items for this particular administration was .83. As it turned out, those values were not too different from the post-equating item parameter estimates. Because of the mean/mean equating, the average of the item parameter estimates were equated to be the same for pre and post equating. Therefore, even when p values tend to be higher with post equating, this is not true with the item difficulty parameter, due to the indeterminacy of the IRT model.

The average absolute difference between the item parameter estimates was 0.31 for Math. The correlation between pre-equating and post operational item parameter estimates was 0.86. Using the criteria mentioned earlier (correlation being 0.90 and average absolute difference being less than 0.20), the item parameter estimates between the two equating models appeared to be somewhat different.

Figure 1 presents raw score -to-theta-scoring tables based on the two equating models mentioned above. The horizontal axis represents the ability estimates, and the vertical axis represents raw scores. According to the figure, the raw score-to-theta scoring tables for pre-equating and post-equating models were very similar, almost overlapping each other throughout the entire score scale. To further observe the impact these different equating models may have, Table 5 was constructed, reporting raw score cuts and percentage of students in each of the performance levels based on the entire testing population in that administration.

As shown in Table 5, the raw score cut corresponding to the scale score for Level II, the passing performance level, was one point higher using the post-equating model as compared to the pre-equating operation model. This resulted in 5.12% more students classified into the below proficient category based on post equating results. The raw score cut corresponding to a scale score of Level III, the advanced performance level, for the post-equating model was one point lower than the pre-equating results, resulting in about 1.57% more students being classified into Level III based on post-equating.

Classification accuracy index based on the two equating models were also computed. Table 6 reports the results. As can be observed from the table, classification accuracy index (sum of the diagonal elements in italics) for pre-equating is 85.5% and is 85.4% for post-equating, very similar results.

Science

Table 3 presents item parameter estimates contrast for Science. Again, as can be observed from this table, the p-values tend to be higher for operational testing. The correlation between pre-equating and post operational item parameter estimates was only 0.69, and the average absolute difference between item parameters for the two equating models was 0.48. It appears that the difference between item parameter estimates was much larger than desired.

Figure 2 presents raw score-to-theta-scoring tables based on the two equating models mentioned above. Again, the horizontal axis refers to the proficiency scale and the vertical axis refers to raw scores. Surprisingly, as can be observed from the figure, even though the item parameter estimates were very different for Science, the raw-to-theta scoring tables for pre and post equating were very similar, almost overlapping each other throughout the entire score scale. Again, one important feature we would like to observe is to look through percentages of students falling into each performance level. Table 5 reports raw score cuts and percent of students.

As can be observed from Table 5, the raw score cuts for both passing and advanced performance levels were the same based on the two equating models, hence producing exactly the same percentage of students in each performance level.

Classification accuracy index based on the two equating models were also computed. Table 7 reports the results. As can be observed from the table, classification accuracy index (sum of the diagonal elements in italics) for pre-equating and post-equating were identical, being 88.5%.

Social Studies

Table 4 presents item parameter estimates contrast for Social studies. Again, the p-values tend to be higher for operational testing. The correlation between pre-equating and post operational item parameter estimates was only 0.73, and the average absolute difference between item parameters for the two equating models was 0.58. It appears that the difference between item parameter estimates was much larger than desired.

Figure 3 presents raw score -to-theta-scoring tables based on the two equating models mentioned above. Compared with Math and Science, Social Studies had the most different scoring tables between the two equating models. The test appeared to be more difficult based on post-equating at lower proficiency level and easier based on post-equating at higher proficiency level. Table 5 also reports the raw score cuts and percentages of students at each performance level for this subject. As can be observed from Table 5, the raw score cut at Level II, the pass level, was two points higher based on pre-equating compared with post-equating. This resulted in

1.23% less students classified into the below proficient category based on the post equating results. The raw score cut corresponding to Level III, the advanced level, was the same for the two equating models.

Classification accuracy index based on the two equating models were also computed. Table 8 reports the results. As can be observed from the table, classification accuracy index (sum of the diagonal elements in italics) for pre-equating and post-equating were the same, being 87.5%.

Discussion

One of the major concerns for pre-equating from the researcher is that the field test is a stand-alone event. Therefore, students' motivation might not be as strong as it would be if the scores are of consequences. There are some constructed response items on some of the tests, with score points ranging from 2 to 6. Especially for those 6-point items, students' responses from field testing tended less than ideal. The omit rate is high and sometimes it is difficult to obtain reasonable item parameters due to relatively large amount of missing data.

Item parameters, scoring tables and student performance classification accuracy were computed based on each of the equating models and compared for three subjects: Math, Science and Social Studies. The following presents the general observations:

1. Post operational results tend to produce higher p-values than the stand-alone field testing, agreeing with the low motivation speculation.
2. Item parameter estimates based on the two equating models tend to be different.
3. Scoring tables and their impact on classification of students tend to be similar between the two equating models.
4. Classification accuracy was similar between two equating models.

Differences on item parameter estimates were observed for all three subjects. However, the scoring tables, especially cut scores, were not as different and sometimes identical (Science) between the two equating models. Such observation is possible because differences at the item level tend to cancel each other out when results are aggregated to score conversion tables. It appeared that the pre-equating model, although facing potentially low student motivation and strong statistical assumptions, seemed to be relatively robust when aggregated into scoring tables.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Joint Technical Committee, (1999). *The Standards for Educational and Psychological Testing.*

Bejar, I.I., & Wingersky, M.S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement, 6*(3), 309-325.

Eignor, D.R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections* (Research report 85-10). Princeton, NJ: Educational Testing Service.

Eignor, D.R., & Stocking, M.L. (1986). *An investigation of the possible causes for the inadequacy of IRT preequating* (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2004*). Test Equating:, Methods and Practices. (2nd ed.)*. New York: Springer-Verlag.

Kolen, M.J., & Harris, D.J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*, 27-39.

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

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment, Research & Evaluation. Vol. 10, Number 13.*

Table 1. Test Layout for the Three Subjects.

	Number of MC Items	Number of CR Items	Number of Essay Items	Sample Size
Math	20	14	N/A	8.295
Science	50	34	N/A	11,128
Social Studies	50	8	2	31,233

Table 2. Contrasts between Pre–equated and Post Operational Item Parameter Estimates, Math

Item	Pre-equated Item Mean	Post Operational Item Mean	Pre-equated Item Parameters	Post Operational Item Parameters	Pre-Post Difference
1	1.60	1.73	-1.53	-1.31	0.22
2	1.56	1.80	-1.39	-1.67	-0.28
3	1.54	1.52	-1.33	-0.53	0.80
4	1.50	1.74	-1.22	-1.39	-0.17
5	1.44	1.81	-1.10	-1.75	-0.65
6	1.44	1.40	-1.05	-0.17	0.88
7	1.38	1.68	-0.90	-1.11	-0.21
8	1.34	1.67	-0.84	-1.03	-0.19
9	1.34	1.46	-0.80	-0.36	0.44
10	1.14	1.48	-0.36	-0.40	-0.04
11	1.04	1.27	-0.15	0.17	0.32
12	1.04	1.42	-0.11	-0.23	-0.12
13	1.04	1.39	-0.10	-0.15	-0.05
14	0.98	1.18	0.01	0.38	0.37
15	0.94	1.38	0.10	-0.13	-0.23
16	0.92	1.25	0.14	0.20	0.06
17	0.86	1.32	0.30	0.02	-0.28
18	0.84	1.33	0.42	0.00	-0.42
19	0.76	1.07	0.55	0.64	0.09
20	0.70	1.05	0.62	0.67	0.05
21	1.18	1.28	-0.38	0.09	0.47
22	1.18	1.71	-0.32	-0.74	-0.42
23	0.74	1.29	0.38	0.33	-0.05
24	0.81	1.20	0.47	0.33	-0.14
25	0.67	1.21	0.64	0.37	-0.27
26	0.54	1.13	0.71	0.58	-0.13
27	1.43	1.15	0.40	1.54	1.14
28	0.78	2.22	1.00	0.63	-0.37
29	1.60	2.41	0.19	0.45	0.26
30	1.12	1.71	0.76	1.07	0.31
31	1.17	2.81	0.87	0.20	-0.67
32	0.86	2.45	0.84	0.48	-0.36
33	0.99	3.19	0.83	0.74	-0.09
34	1.28	3.23	0.83	0.70	-0.13

Table 3. Contrasts between Pre–equated and Post Operational Item Parameter Estimates, Science.

Item	Pre-equated Item Mean	Post Operational Item Mean	Pre-equated Item Parameters	Post Operational Item Parameters	Pre-Post Difference
1	0.86	0.93	-2.00	-1.73	-0.27
2	0.66	0.81	-0.60	-0.47	-0.13
3	0.60	0.90	-0.28	-1.33	1.05
4	0.39	0.64	0.76	0.62	0.14
5	0.37	0.50	0.89	1.33	-0.44
6	0.56	0.70	-0.07	0.28	-0.35
7	0.62	0.71	-0.38	0.18	-0.56
8	0.67	0.78	-0.64	-0.21	-0.43
9	0.63	0.76	-0.43	-0.09	-0.34
10	0.46	0.59	0.44	0.87	-0.43
11	0.46	0.64	0.43	0.58	-0.15
12	0.54	0.71	0.03	0.18	-0.15
13	0.70	0.83	-0.85	-0.58	-0.27
14	0.89	0.94	-2.22	-1.94	-0.28
15	0.51	0.73	0.18	0.09	0.09
16	0.79	0.88	-1.37	-1.13	-0.24
17	0.51	0.74	0.15	0.01	0.14
18	0.60	0.73	-0.25	0.09	-0.34
19	0.26	0.39	1.52	1.91	-0.39
20	0.62	0.83	-0.39	-0.59	0.20
21	0.56	0.67	-0.06	0.46	-0.52
22	0.56	0.66	-0.05	0.48	-0.53
23	0.52	0.66	0.10	0.50	-0.40
24	0.60	0.72	-0.28	0.16	-0.44
25	0.49	0.60	0.26	0.80	-0.54
26	0.61	0.73	-0.34	0.09	-0.43
27	0.59	0.84	-0.25	-0.68	0.43
28	0.68	0.83	-0.75	-0.61	-0.14
29	0.60	0.82	-0.27	-0.54	0.27
30	0.68	0.87	-0.71	-0.97	0.26
31	0.74	0.76	-1.05	-0.10	-0.95
32	0.84	0.68	-1.74	0.38	-2.12
33	0.34	0.87	1.05	-0.95	2.00
34	0.52	0.76	0.10	-0.10	0.20
35	0.70	0.87	-0.83	-0.93	0.10
36	0.67	0.72	-0.63	0.12	-0.75
37	0.65	0.89	-0.55	-1.24	0.69
38	0.64	0.84	-0.50	-0.67	0.17
39	0.84	0.57	-1.79	0.95	-2.74

40	0.63	0.67	-0.43	0.44	-0.87
41	0.33	0.93	1.13	-1.78	2.91
42	0.50	0.84	0.22	-0.69	0.91
43	0.84	0.57	-1.81	0.99	-2.80
44	0.67	0.63	-0.63	0.68	-1.31
45	0.32	0.77	1.18	-0.21	1.39
46	0.56	0.60	-0.09	0.80	-0.89
47	0.41	0.71	0.68	0.23	0.45
48	0.40	0.49	0.72	1.37	-0.65
49	0.48	0.63	0.34	0.64	-0.30
50	0.50	0.42	0.24	1.76	-1.52
51	0.74	0.82	-1.05	-0.51	-0.54
52	0.46	0.65	0.44	0.58	-0.14
53	0.69	0.83	-0.76	-0.55	-0.21
54	0.58	0.81	-0.16	-0.40	0.24
55	0.42	0.74	0.61	0.08	0.53
56	0.57	0.81	-0.16	-0.40	0.24
57	0.47	0.69	0.38	0.37	0.01
58	0.45	0.66	0.48	0.57	-0.09
59	0.79	0.85	-1.39	-0.72	-0.67
60	0.13	0.33	2.56	2.28	0.28
61	0.16	0.34	2.21	2.24	-0.03
62	0.47	0.74	0.36	0.10	0.26
63	0.47	0.78	0.40	-0.15	0.55
64	0.34	0.56	1.08	1.08	0.00
65	0.37	0.64	0.89	0.68	0.21
66	0.18	0.80	2.13	-0.34	2.47
67	0.68	0.92	-0.68	-1.59	0.91
68	0.42	0.61	0.64	0.81	-0.17
69	0.64	0.90	-0.47	-1.24	0.77
70	0.57	1.34	1.18	0.62	0.56
71	0.53	0.89	0.06	-1.16	1.22
72	0.40	0.83	0.73	-0.55	1.28
73	0.69	0.93	-0.77	-1.65	0.88
74	0.13	0.85	2.54	-0.69	3.23
75	0.78	0.42	-1.31	1.78	-3.09
76	0.64	0.84	-0.48	-0.66	0.18
77	0.18	0.38	2.04	1.99	0.05
78	0.51	0.69	0.17	0.38	-0.21
79	0.48	0.67	0.34	0.47	-0.13
80	0.15	0.37	2.40	2.06	0.34
81	0.16	0.44	2.29	1.72	0.57
82	0.43	0.44	0.60	1.69	-1.09
83	0.39	0.65	0.79	0.58	0.21
84	0.20	0.83	1.90	-0.52	2.42

Table 4. Contrasts between Pre–equated and Post Operational Item Parameter Estimates, Social Studies.

Item	Pre-equated Item Mean	Post Operational Item Mean	Pre-equated Item Parameters	Post Operational Item Parameters	Pre-Post Difference
1	0.81	0.79	-1.18	-0.19	0.99
2	0.69	0.72	-0.37	0.28	0.65
3	0.62	0.71	-0.03	0.33	0.36
4	0.60	0.65	0.07	0.66	0.59
5	0.83	0.85	-1.28	-0.70	0.58
6	0.56	0.67	0.27	0.58	0.31
7	0.67	0.74	-0.28	0.15	0.43
8	0.70	0.77	-0.45	-0.02	0.43
9	0.71	0.75	-0.51	0.10	0.61
10	0.68	0.72	-0.32	0.26	0.58
11	0.71	0.68	-0.48	0.52	1.00
12	0.66	0.81	-0.22	-0.34	-0.12
13	0.50	0.63	0.57	0.80	0.23
14	0.61	0.73	0.04	0.21	0.17
15	0.77	0.83	-0.85	-0.50	0.35
16	0.72	0.76	-0.53	0.05	0.58
17	0.48	0.52	0.66	1.34	0.68
18	0.72	0.71	-0.53	0.35	0.88
19	0.62	0.67	0.00	0.56	0.56
20	0.70	0.81	-0.45	-0.31	0.14
21	0.46	0.57	0.78	1.11	0.33
22	0.70	0.72	-0.45	0.29	0.74
23	0.59	0.68	0.13	0.50	0.37
24	0.68	0.85	-0.34	-0.64	-0.30
25	0.54	0.64	0.40	0.71	0.31
26	0.56	0.60	0.31	0.94	0.63
27	0.55	0.62	0.32	0.87	0.55
28	0.66	0.71	-0.20	0.35	0.55
29	0.55	0.62	0.32	0.83	0.51
30	0.77	0.80	-0.85	-0.26	0.59
31	0.74	0.77	-0.67	-0.04	0.63
32	0.74	0.74	-0.66	0.17	0.83
33	0.74	0.73	-0.68	0.22	0.90
34	0.64	0.68	-0.13	0.51	0.64
35	0.74	0.72	-0.70	0.27	0.97
36	0.84	0.85	-1.36	-0.69	0.67
37	0.58	0.59	0.20	1.02	0.82
38	0.81	0.84	-1.12	-0.60	0.52

39	0.61	0.56	0.03	1.13	1.10
40	0.73	0.72	-0.60	0.29	0.89
41	0.70	0.71	-0.44	0.31	0.75
42	0.60	0.66	0.08	0.60	0.52
43	0.66	0.75	-0.19	0.08	0.27
44	0.67	0.70	-0.28	0.40	0.68
45	0.58	0.62	0.19	0.86	0.67
46	0.58	0.62	0.18	0.84	0.66
47	0.58	0.68	0.20	0.53	0.33
48	0.53	0.63	0.46	0.80	0.34
49	0.77	0.81	-0.89	-0.34	0.55
50	0.53	0.62	0.42	0.82	0.40
51	4.65	7.51	2.26	1.57	-0.69
52	1.14	1.58	0.48	-0.31	-0.79
53	1.45	1.75	-0.32	-0.69	-0.37
54	1.59	1.84	-0.68	-1.18	-0.50
55	0.89	0.88	-1.53	-0.97	0.56
56	1.70	1.85	-0.89	-1.07	-0.18
57	0.77	0.95	-0.56	-1.89	-1.33
58	1.56	1.86	-0.39	-1.06	-0.67
59	0.48	0.70	0.96	0.45	-0.51
60	5.58	8.52	1.93	1.22	-0.71

Table 5. Comparisons of Raw Score Cuts and Percentages of Students in Each of the Performance Levels between Pre-equating and Post Equating Models.

	Pre–Equating Model		Post Equating Model	
Scale Score	**Raw Score Cut**	**Percent**	**Raw Score Cut**	**Percent**
Math (correlation = 0.86; average abs diff = 0.32)				
Level I		28.43		33.55
Level II	47	47.61	48	40.93
Level III	71	23.95	70	25.52
Science (correlation = 0.69; average abs diff = 0.48)				
Level I		24.60		24.60
Level II	48	52.71	48	52.71
Level III	73	22.69	73	22.69
Social Studies (correlation=0.73; average abs diff=0.58)				
Level I		30.44		29.21
Level II	56	36.36	54	37.59
Level III	71	33.20	71	33.20

Table 6. Classification Accuracy Index for Math[1].

Level	LEVEL 1	LEVEL 2	LEVEL 3	True
Pre-equating				
LEVEL 1	*26.0*	6.0	0.0	31.9
LEVEL 2	1.6	*39.2*	3.8	44.7
LEVEL 3	0.0	2.8	*20.3*	23.0
Expected	27.6	48.0	24.1	99.6
Post-equating				
LEVEL 1	*28.9*	4.7	0.0	33.6
LEVEL 2	3.1	*34.4*	3.6	41.2
LEVEL 3	0.0	2.8	*22.1*	24.9
Expected	32.1	41.9	25.7	99.6

Table 7. Classification Accuracy Index for Science.

Level	LEVEL 1	LEVEL 2	LEVEL 3	True
Pre-equating				
LEVEL 1	*23.2*	4.1	0.0	27.3
LEVEL 2	1.4	*46.5*	3.4	51.3
LEVEL 3	0.0	2.1	*18.8*	20.9
Expected	24.6	52.7	22.2	99.5
Post-equating				
LEVEL 1	*23.2*	4.1	0.0	27.3
LEVEL 2	1.4	*46.5*	3.4	51.3
LEVEL 3	0.0	2.1	*18.8*	20.8
Expected	24.6	52.7	22.2	99.5

[1] Due to the calculation and the use of +10 and -10 as cut offs at the two extremes, the overall sum of true and observed was not always 100%, but it should be very close to 100%.

Table 8. Classification Accuracy Index for Social Studies.

Level	LEVEL 1	LEVEL 2	LEVEL 3	True
Pre-equating				
LEVEL 1	*29.4*	4.5	0.0	33.9
LEVEL 2	1.6	*27.7*	3.5	32.8
LEVEL 3	0.0	2.3	*30.4*	32.8
Expected	31.0	34.6	33.9	99.5
Post-equating				
LEVEL 1	*27.1*	3.5	0.0	30.6
LEVEL 2	2.0	*30.3*	3.9	36.2
LEVEL 3	0.0	2.7	*30.1*	32.7
Expected	29.2	36.4	33.9	99.5

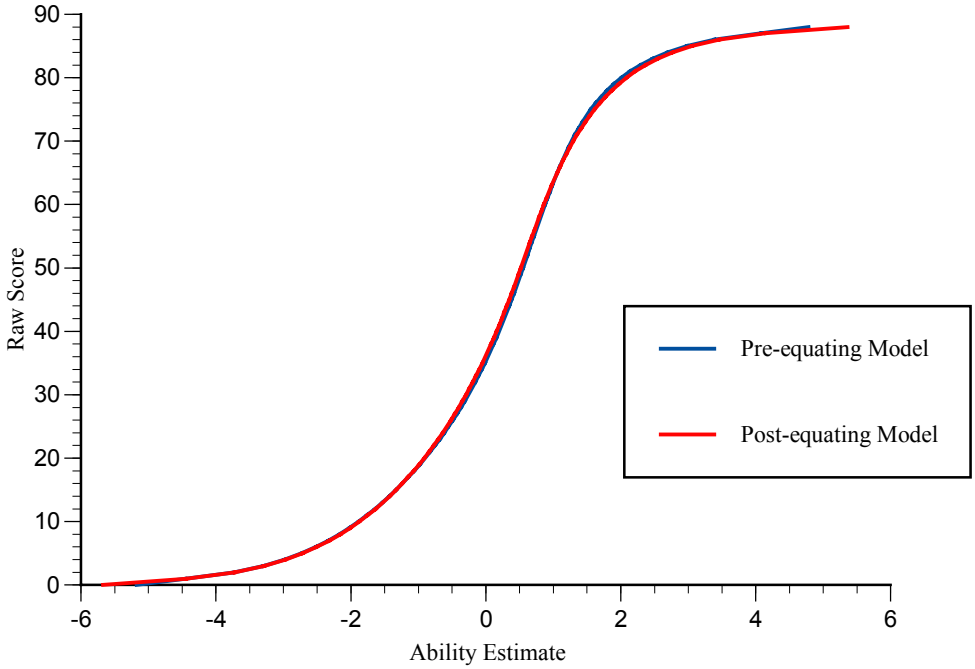

Figure 1. Comparison of Relationship between Raw Score and Ability Estimates between Pre-equating Model and Post-equating Model, Math.

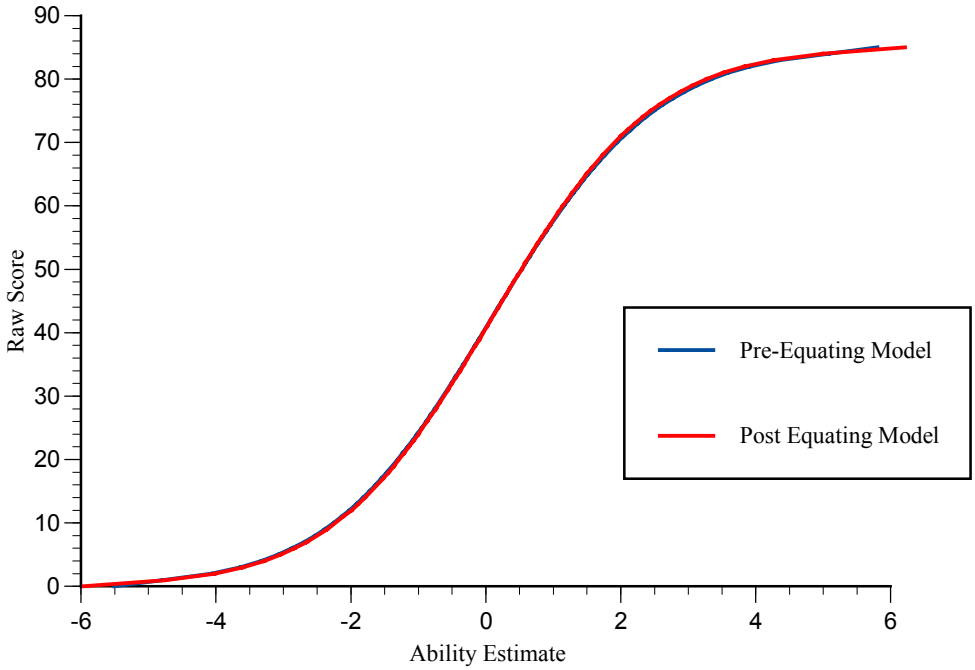

Figure 2. Comparison of Relationship between Raw Score and Ability Estimates between Pre-equating Model and Post-equating Model, Science

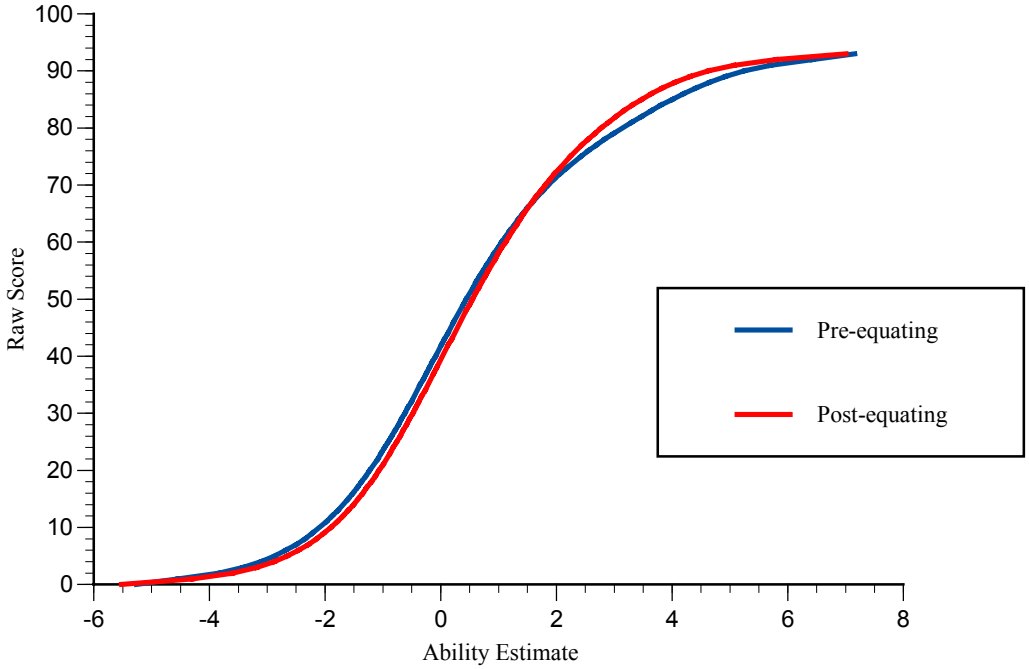

Figure 3. Comparison of Relationship between Raw Score and Ability Estimates between Pre-equating Model and Post-equating Model, Social Studies.