Listening. Learning. Leading.®

Research Report
ETS RR–12-12

Does Preequating Work? An Investigation Into a Preequated Testlet-Based College Placement Exam Using Postadministration Data

Rui Gao

Wei He

Chunyi Ruan

July 2012

Does Preequating Work?
An Investigation Into a Preequated Testlet-Based College Placement Exam Using Postadministration Data

Rui Gao
ETS, Princeton, New Jersey

Wei He
Northwest Evaluation Association (NWEA), Portland, Oregon

Chunyi Ruan
ETS, Princeton, New Jersey

July 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: James E. Carlson

Technical Reviewers: Yanmei Li and Robert L. Smith

Abstract

In this study, we investigated whether preequating results agree with equating results that are based on observed operational data (postequating) for a college placement program. Specifically, we examined the degree to which item response theory (IRT) true score preequating results agreed with those from IRT true score postequating and from observed score equating. Three academic subjects were examined in this study: analyzing and interpreting literature, American government, and college algebra. The findings suggested that differences between equating results from IRT true score preequating and postequating varied from subject to subject. In general, IRT true score postequating agreed with IRT true score preequating for most of the forms for a test subject. Any difference among the equating results can be attributed to the way through which items were pretested, contextual/order effects, or the violation of IRT assumptions.

Key words: IRT preequating, IRT true score equating, observed score equating, IRT postequating

Equating is a statistical process used to adjust scores on two or more forms of a test so that the scores can be used interchangeably (Kolen & Brennan, 2004). Depending on when equating is conducted, it can be further categorized as preequating and postequating. Preequating, as the name implies, refers to the process through which conversions from raw to scaled scores are established prior to the time the new test is administered operationally as an intact final form. Preequating often is based on item response theory (IRT).

Due to the fact that preequating can establish the conversion table prior to operational testing, there are a number of advantages in using preequating rather than postequating (see Eignor, 1985; Kirkpatrick & Way, 2008; and Kolen & Brennan, 2004, for a complete review). These advantages include more flexible assessment and a better quality control check for the tests. Perhaps the most appealing feature of preequating is the ability to facilitate immediate score reporting for tests that require reporting scores immediately after the test administration. The *College-Level Examination Program*® (*CLEP*®) or *Graduate Record Examinations*® (*GRE*®) tests are good examples. Due to these advantages, preequating is sometimes used in large-scale assessments.

Much research has been conducted to compare the difference in equating results between preequating and postequating. In general, results have not been consistent. Eignor (1985), in investigating the feasibility and practical outcomes of preequating *SAT*® verbal and mathematical sections through IRT true score equating, found that preequating worked adequately for the verbal sections but not for the mathematical section. Similarly, Kolen and Harris (1990), in comparing item preequating and random groups equating using IRT and equipercentile methods, found that preequating performed poorly for the ACT math test. Contributing reasons for poor preequating results mainly focused on the inconsistent behaviors of test items in pretest and operational contexts. These inconsistencies can be caused by a lack of motivation on the part of test-takers for answering pretest items, different ability distributions used for pre- and postequating, change in item position, context effects, violation of assumptions underlying the use of IRT models, and model-data misfit (Eignor & Stocking, 1986; Kolen & Harris, 1990; Stocking & Eignor, 1986; Tong, Wu, & Xu, 2008). However, other studies suggested that preequating can achieve satisfactory results. Bejar and Wingersky (1982) compared the preequating results of the Test of Standard Written English (TSWE) using different equating methods and concluded that preequating could be a feasible operational procedure. Livingston

(1985), using a method similar to regression, demonstrated that preequating was highly accurate in three of the four New Jersey College Basic Skills Placement tests. Studies conducted by Domaleski (2006) and Tong et al. (2008) resulted in similar pre- and postequated scoring tables and similar accuracy of classifying students into different performance levels, thus supporting the use of preequating.

Apart from different research findings for preequating, a literature review indicates that little research has been conducted on whether preequating agrees with postequating for a testlet-based and computer-administered testing program. To this end, this study used real data to investigate whether preequating results agree with equating results based on operational data (postequating) for this type of test. The study examined the degree to which IRT true score preequating results agreed with those obtained from IRT true score postequating and observed score equating methods.

Method

Introduction to the Testing Program

Exams for the testing program in this study are administered in computer-based testing (CBT) format. Testlets are the building blocks for the exams. A testlet is a collection of questions from a coherent content domain. Multiple versions of parallel testlets are created for each content domain; these are combined, one testlet from each content domain, to build unique but parallel forms in terms of content and statistical properties. For example, the exam for Analyzing and Interpreting Literature (AIL) consists of three types of operational testlets, A, B, and C, from three different content domains. Each type of testlet has two parallel versions (A1, A2, B1, B2, C1, and C2). Therefore, different combinations of testlets from each content domain result in eight parallel forms: A1B1C1, A1B1C2, A1B2C1, A1B2C2, A2B1C1, A2B1C2, A2B2C1, and A2B2C2.

Pretest items are embedded in operational testlets. The testlets and pretest items assembled for each administration are called a package. Each package is administered continuously in the field for approximately 3 years until a new package is assembled by replacing a portion of items in the old package with pretest items. This occurs when enough data are accumulated on the items (500 responses for each pretest item). In this study, the 2001 and 2004 packages were used.

For AIL, these eight test forms overlap with one another to varying degrees at the testlet level. The computerized delivery software assigns a test form at random to a test-taker. Test scores on different forms of the 2004 package are equated to the common reference form, A2B2C2 in the 2001 package, to adjust for form differences.

The Rasch model is used for item calibration, and IRT true score preequating is used operationally to report scale scores, which typically range from 20 to 80. To derive a raw-to-scale score conversion table for each form of a new package, the following procedures are generally followed:

$$\text{Number-Correct Score}_{new} \rightarrow \theta_{new} \rightarrow \theta_{reference} \rightarrow \text{Number-Correct Score}_{reference} \rightarrow \text{CLEP Scaled Score}$$

To be specific, for a particular new form, the observed number-correct scores on the form are treated as expected IRT true scores, which are then converted to the ability scale (θ_{new}) based on the Rasch model. The Stocking and Lord (1983) transformation method is used to place all parameter estimates on the same metric. Then, ability scores (θ_{new}) on the scale of reference form are converted to expected IRT true scores, which are then treated as if they are reference-form number-correct scores. Finally, using a linear conversion associated with the reference form, these raw scores on the reference-form scale are placed onto the 20-to-80 score scale.

Data

To allow the comparison of the difference in equating results between pre- and postequating (i.e., equating based on the postadministration data for the different forms of test in the 2004 packages), data obtained from two different packages were used in this study: (a) the 2001 package, which contained data collected from 2001 to 2004 on the pretest items that were used in the 2004 package as operational items, and (b) the 2004 package, which contained operational data collected from 2004 to 2008.

Three different subjects were used in this study: analyzing and interpreting literature (AIL), American government (GOV), and college algebra (ALG). For AIL, eight forms were used for this study, and each form had data from about 3,500 examinees; for GOV, seven forms were used, and each form had about 1,700 examinees; for ALG, eight forms were used, and each form had about 850 examinees.

Equating Design and Equating Methods

Preequating. IRT true score equating was used for preequating. To preequate the 2004 package for each subject, the operational item response data collected from 2001 to 2004 on the 2001 package were calibrated first. Then, both the pretest and operational item response data were calibrated by fixing the parameters of the operational items. The resulting parameters were used to preequate the forms in the 2004 package to the reference form through IRT true score equating. The raw-to-scale score conversion table was created for each form.

Postequating. IRT true score equating and observed score equating methods were used for postequating.

IRT true score equating. To conduct IRT true score postequating, operational data collected from 2004 to 2008 on the 2004 package were calibrated. These item parameters were then transformed to the base scale using the Stocking-Lord method with common items between the 2004 and 2001 packages as an anchor. The resulting parameters were then used to postequate the forms in the 2004 package to the reference form. The raw-to-scale score conversion table was created for each form.

Observed score equating. The observed score equating methods for postequating were based on either an equivalent groups without anchor items (EG) or a nonequivalent groups with anchor test (NEAT) design.

The EG design was used when there were no common items between the new and reference forms. Because examinees were administered a randomly selected test form, it was reasonable to assume that examinees who took tests from different packages were equivalent.

Group equivalence was examined to decide upon which design to use for new forms sharing common items with the reference form. The standardized difference in mean anchor scores between the new and reference groups was computed to examine group equivalence. For all forms involved in this study, their standardized differences were below .1, and the majority of the differences were below .05, indicating the groups were regarded as equivalent, according to Cohen (1988). The EG design was used for AIL and GOV. For ALG, except for the two forms having no common items with the reference form, due to the relatively small sample sizes (80-850), the NEAT design was used to equate the remaining six forms because the use of an anchor can reduce the standard error of equating (Kolen & Brennan, 2004).

Both the equipercentile and mean-sigma linear equating methods were used to equate each form using EG design. To decide upon which conversion table to use as the result for observed score equating, the concept of the difference that matters (DTM; Dorans & Feigenbaum, 1994; Dorans, Holland, Thayer, & Tateneni, 2003) was adopted to evaluate the magnitude of the difference between the two conversion tables. Generally, if the difference between linear and equipercentile equating was within the range specified by DTM, the conversion table from the linear equating method was adopted. Otherwise, the conversion table from the equipercentile equating was used. A difference of .5 was considered as significant because it resulted in a change in the reporting score. Based on the examination of DTM, conversion tables produced from equipercentile equating were used for AIL forms and two ALG forms; those from mean-sigma equating were used for all GOV forms.

For the six ALG forms using the NEAT design, both the nonlinear chained equipercentile and the linear Tucker equating methods were used to equate each form. Similar to the EG design, DTM was adopted to decide which conversion table to use. Conversion tables from chained equipercentile were used for three forms, and those from Tucker equating were used for the other three.

Evaluation Criteria

The results from the IRT true score preequating method were used as the reference. The conversion lines yielded from different postequating methods were compared against the conversion line from IRT true score preequating. DTM was again employed to compare the difference in equated scores between the postequating and preequating methods.

Pass/fail classification rates given by different equating methods were also reported. Test scores are reported on a 20-to-80 scale. Each test has two cut scores, C and B cuts. A scaled score of 50 is used as the C-cut score across all test titles. The B-cut score varies across different test titles. Classification rates for both C and B cuts were reported for each of the three tests in the study.

In addition, three other indices were employed: mean signed difference (MSD; Equation 1), root mean square difference (RMSD; Equation 2), and mean absolute difference (MAD; Equation 3). All three indices are weighted by the frequency at number-correct raw score.

$$\text{MSD} = \frac{\sum_i f_i (X_i^{'} - X_i)}{\sum_i f_i}, \quad (1)$$

$$\text{RMSD} = \sqrt{\frac{\sum_i f_i (X_i^{'} - X_i)^2}{\sum_i f_i}}, \quad (2)$$

$$\text{MAD} = \frac{\sum_i f_i \mid X_i^{'} - X_i \mid}{\sum_i f_i}, \quad (3)$$

where f_i is the frequency at number-correct raw score level *i*, $X_i^{'}$ is the equated score at number-correct raw score level *i*, and X_i is the equated score from IRT true score preequating at number-correct raw score level *i*. All three indices summarize the difference between a pair of conversions over score points. The MSD and MAD indices are measures of the mean difference between converted scores; they are also called *bias indices*. MSD shows the direction of the difference, while MAD shows the absolute magnitude of the difference. RMSD is a measure of the mean squared difference between equated scores. It provides an index of similarity of conversions that is weighted by score frequency (Kolen & Harris, 1990).

Results

Results for different test subjects are presented separately in this section. It should be noted that, for all forms, the results from the IRT true score preequating method were used as the baseline for comparison. Approximately 50 examinees at the top and bottom score scale were excluded from all plots. The reason is that the equating results at the ends of the scale are not stable due to the small sample sizes, thus adding noise in the comparison of different conversion lines. Excluding them from the plot can present a clearer picture of the pattern of the conversion lines. In the plots, the number-correct raw scores corresponding to the C cut for each form are indicated by the vertical dotted lines. The number-correct raw score corresponding to the B cut is indicated by the vertical solid lines. The DTM bands are indicated by the two horizontal dotted lines.

Analyzing and Interpreting Literature

For the eight AIL forms, the standardized differences in mean anchor scores between the reference and new groups were all below .08. The differences in equated scores between

equipercentile and mean-sigma methods exceeded DTM at a number of raw score levels. Therefore, the conversion tables obtained from the equipercentile equating using the equivalent group design (Eq%_EG) were considered appropriate to use as the results from the observed score equating method.

For the sake of space limits, conversion lines are selectively presented. For all eight AIL forms, IRT true score preequating consistently yielded the highest scale score at the score range below the C-cut score, followed by IRT true score postequating and the equipercentile method. That said, with the IRT true score preequating method, the examinees in this range would gain up to 3 more scale score points than with the equipercentile method. This difference suggests that IRT true score preequating made the test appear harder than it actually was for the examinees with raw scores below the C cut. Figure 1, which depicts the conversion line for Form 1, can shed light on the above finding.

As raw scores increase above the C-cut score, the difference between the scaled scores yielded by IRT true score preequating and the two postequating methods becomes smaller. The scaled scores yielded by IRT true score preequating become slightly lower than those yielded by the postequating methods at some point above this cut. However, except for Form 2, whose conversion line is presented in Figure 2, the difference falls within the DTM bands for most scores above this cut.

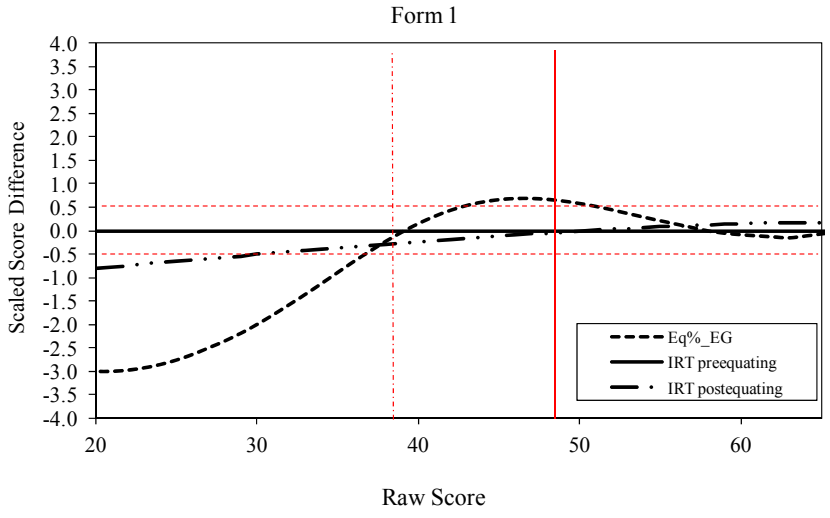

***Figure 1*. Scaled score difference for analyzing and interpreting literature (AIL) Form 1. Eq%_EG = equipercentile equating using the equivalent group design; IRT = item response theory.**

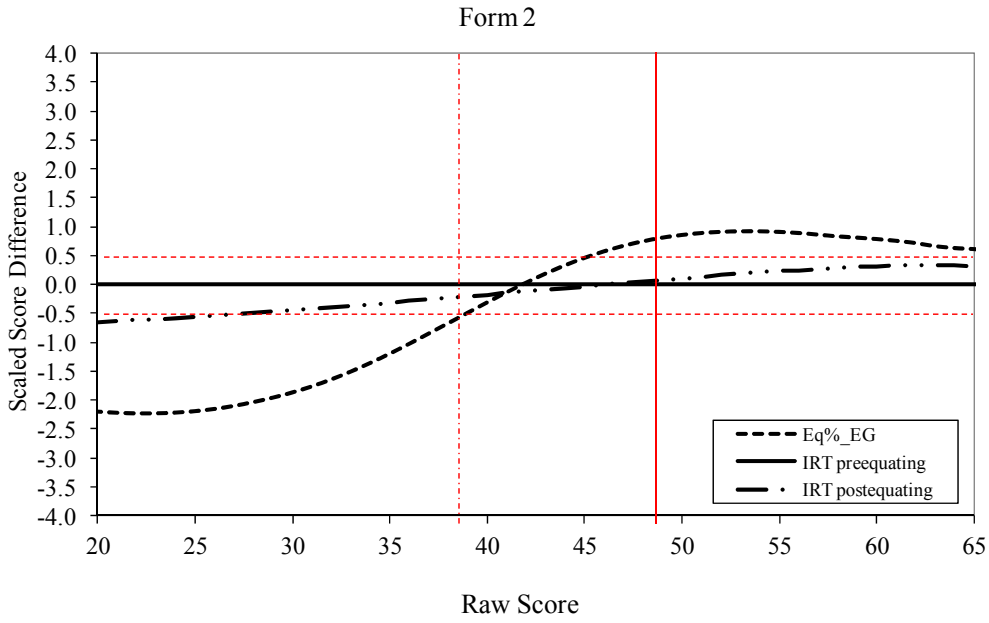

Figure 2. **Scaled score difference for analyzing and interpreting literature (AIL) Form 2. Eq%_EG = equipercentile equating using the equivalent group design; IRT = item response theory.**

The above finding, in general, echoed with the descriptive statistics of equated scores from different equating methods. That is, IRT true score preequating tends to yield a higher average scaled score (58.0) consistently than those from the equipercentile equating (57.7) and IRT true score postequating (57.6).

Regarding the classification rate, IRT true score preequating tends to result in more examinees passing the C cut than any other two postequating methods and, in general, equipercentile equating tends to result in the lowest passing rate. On average, the passing rate based on equipercentile equating is 29.1%, 31.8% based on IRT true score preequating, and 31.0% based on IRT true score postequating.

Table 1 presents the three indices used to evaluate the equating results with IRT true score preequating results as the baseline. All three indices indicate that IRT true score postequating yields results closer to the IRT true score preequating method, having smaller RMSD, MSD, and MAD in all forms except for Forms 7 and 8.

Table 1

RMSD*, *MSD*, *and MAD for Analyzing and Interpreting Literature (AIL)

	RMSD		MSD		MAD	
Form	Obs.	IRT post	Obs.	IRT post	Obs.	IRT post
1	1.141	0.326	-0.345	-0.208	0.816	0.253
2	1.099	0.290	-0.208	-0.102	0.927	0.239
3	0.920	0.779	-0.171	-0.546	0.715	0.615
4	0.882	0.696	-0.353	-0.400	0.669	0.553
5	1.217	0.545	-0.580	-0.452	0.884	0.463
6	1.011	0.464	-0.536	-0.330	0.744	0.383
7	0.824	0.997	-0.444	-0.785	0.602	0.818
8	0.536	0.915	-0.171	-0.661	0.448	0.746

Note. IRT post = IRT true score postequating; MAD = mean absolute difference; MSD = mean signed difference; Obs = observed score equating; RMSD = root mean square difference.

American Government

For the seven GOV forms, the standardized differences in mean anchor scores between the reference and new groups are all below .02. The differences in equated scores between equipercentile and mean-sigma methods are within DTM bands for a majority of raw score levels. Therefore, the conversion tables obtained from the mean-sigma equating method using the equivalent group design (MS_EG) was considered appropriate to use as the results from the observed score equating method.

There were four GOV forms for which IRT true score preequating yielded slightly lower scaled scores than the IRT true score postequating method. The differences between the two equating methods slightly exceed the DTM bands, suggesting that IRT true score preequating tends to make the test appear slightly easier than the IRT true score postequating method. For example, the conversion line for Form 1 is presented in Figure 3. The rest of the three forms did not see any significant difference between the scaled scores from IRT true score preequating and IRT true score postequating.

IRT true score preequating appears to yield higher scaled scores than the observed score equating method and made the test appear slightly harder, although the patterns of the differences between the scaled scores yielded by the two equating methods are inconsistent across different forms. In one form, Form 1, whose conversion line is portrayed in Figure 3, saw little difference. In four forms, with Form 4 as an example, indicated in Figure 4, IRT true score preequating yields higher scaled scores at most of the raw score levels. In two forms, with

Form 6 as an example, presented in Figure 5, IRT true score preequating consistently yields higher scaled scores.

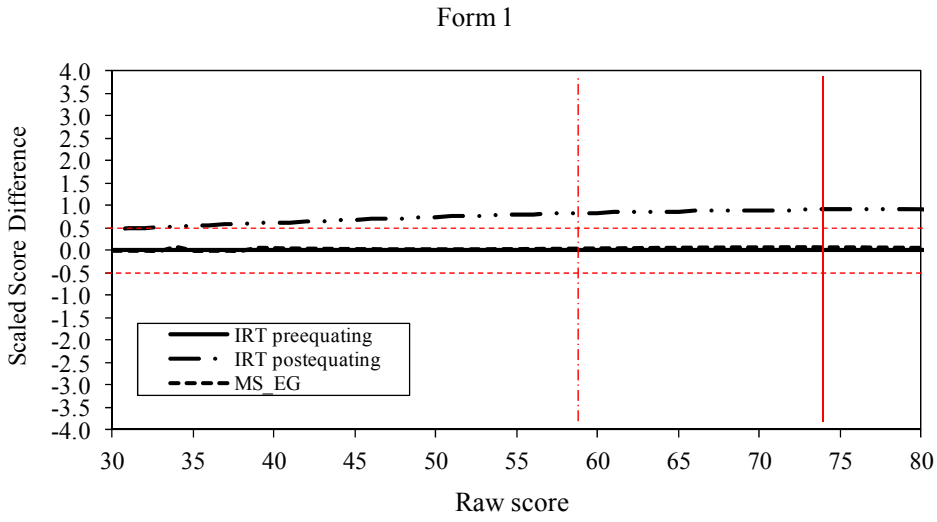

Figure 3. **Scaled score difference for American government (GOV) Form 1. MS_EG = mean-sigma equating method using the equivalent group design; IRT = item response theory.**

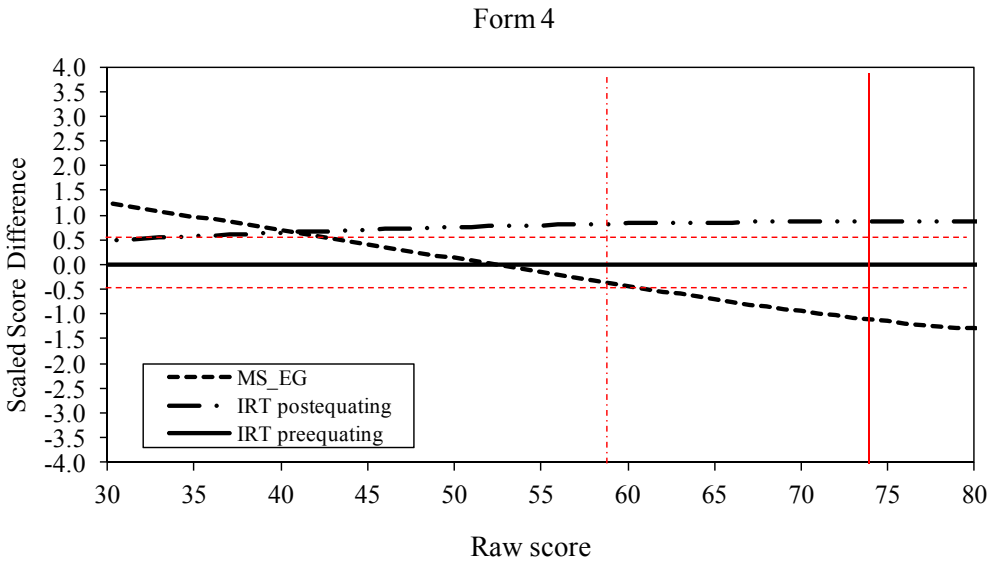

Figure 4. **Scaled score difference for American government (GOV) Form 4. IRT = item response theory; MS_EG = mean-sigma equating method using the equivalent group design.**

.

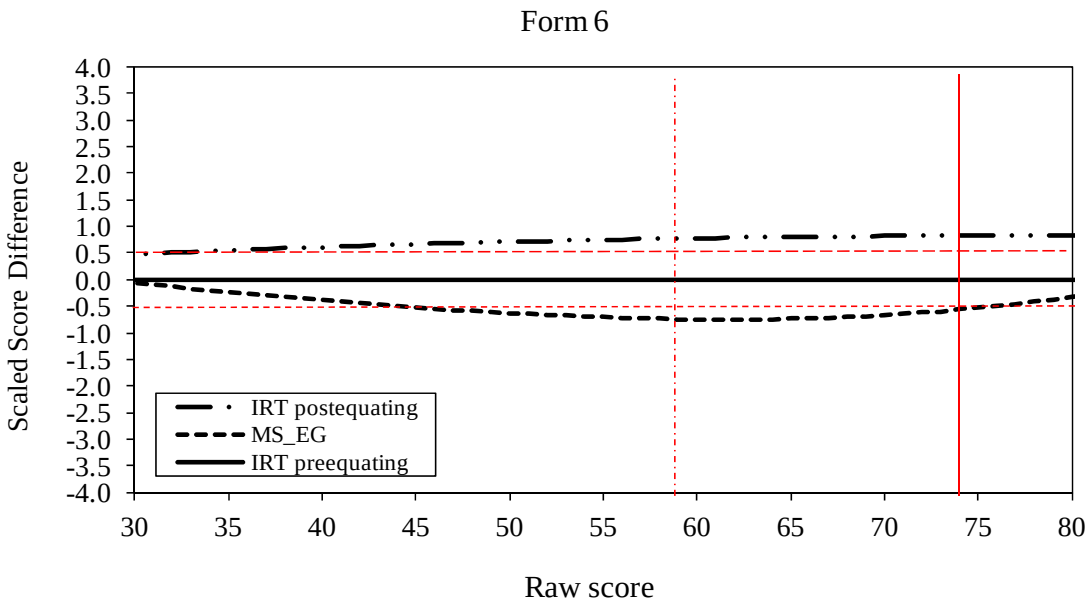

Figure 5. **Scaled score difference for American government (GOV) Form 6. MS_EG = mean-sigma equating method using the equivalent group design.**

For all GOV forms, IRT true score preequating method tends to result in fewer examinees passing (40.4%) than the IRT true score postequating method (41.9%). The IRT true score preequating yielded slightly higher passing rates than the observed score equating method (39.9%).

The examination of the descriptive statistics found that IRT true score preequating yields slightly lower average scaled scores (49.6) than the IRT true score postequating method (50.0). The observed score equating method tends to yield the lowest average scaled scores (49.1) consistently.

Table 2 presents the three indices used to evaluate the equating results with IRT true score preequating results as the baseline. Values within $\pm$ .1 are in boldface. All three indices indicated that, for Forms 3, 5, and 7, IRT true score postequating yields results close to those for IRT true score preequating. For Forms 1, 2, 4, and 6, the indices suggest that the results from IRT true score preequating differ from those from the two postequating methods in different directions.

Table 2

RMSD, MSD, and MAD for American Government (GOV)

	RMSD		MSD		MAD	
Form	Obs.	IRT post	Obs.	IRT post	Obs.	IRT post
1	0.056	0.803	0.000	0.794	0.041	0.794
2	0.400	0.735	-0.169	0.745	0.348	0.745
3	0.725	0.040	-0.651	-0.038	0.669	0.038
4	0.784	0.795	-0.308	0.787	0.664	0.787
5	0.934	0.021	-0.585	-0.006	0.783	0.018
6	0.620	0.748	-0.587	0.741	0.594	0.741
7	0.823	**0.044**	-0.803	**-0.043**	0.803	**0.043**

Note. Values within + .1 are in boldface. IRT post = item response theory true score postequating; MAD = mean absolute difference; MSD = mean signed difference; Obs. = observed score equating; RMSD = root mean square difference.

Algebra

For two ALG forms that did not have common items with the reference form, an examination of DTM found the conversion table from Eq%_EG appropriate to use. Both Tucker and chained-equipercentile equating (Eq%_NEAT) were conducted on the remaining six forms. For three forms having differences in equated scores obtained from those two methods within the DTM bands, the conversion tables obtained from the Tucker method were used. For the rest of the three forms having the differences exceeding DTM bands for a number of raw score levels, the conversion tables obtained from Eq%_NEAT were used.

IRT true score preequating generally yields lower scaled scores than the two postequating methods, suggesting that preequating tends to make the test appear easier than it actually was. The differences between the results from IRT true score preequating and IRT true score postequating are smaller than those between IRT true score preequating and observed score equating.

Except for Form 1, whose conversion line is presented in Figure 6, the differences between the results from IRT true score preequating and IRT true score postequating are roughly within the DTM for scaled scores higher than the C cut. Speaking of the differences between the results from IRT true score preequating and observed score equating, there are three forms having differences within the DTM for scaled scores higher than the C cut. An example is

provided for Form 5 in Figure 7. There are another five forms having differences beyond the DTM, and an example can be referred to in the conversion line for Form 1 in Figure 6.

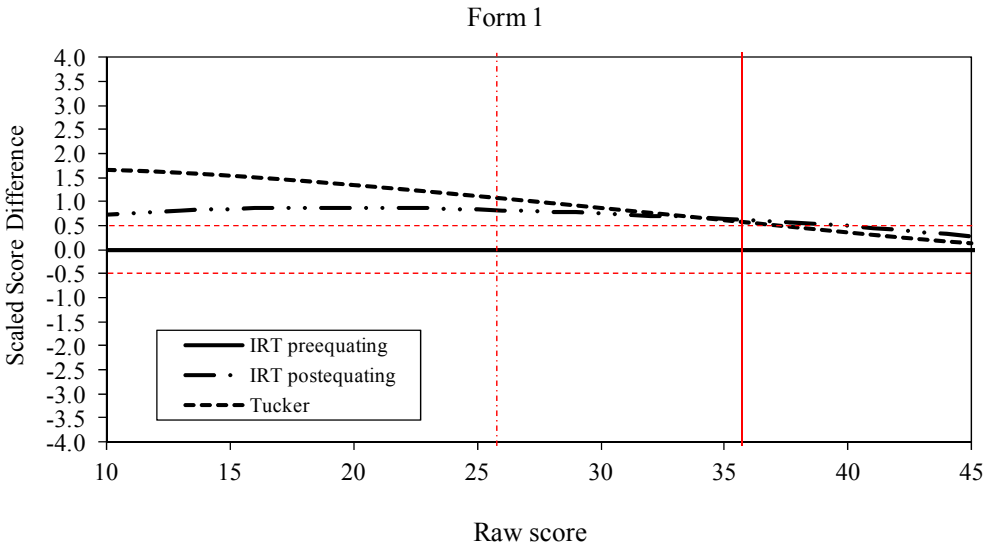

Figure 6. **Scaled score difference for college algebra (ALG) Form 1. IRT = item response theory.**

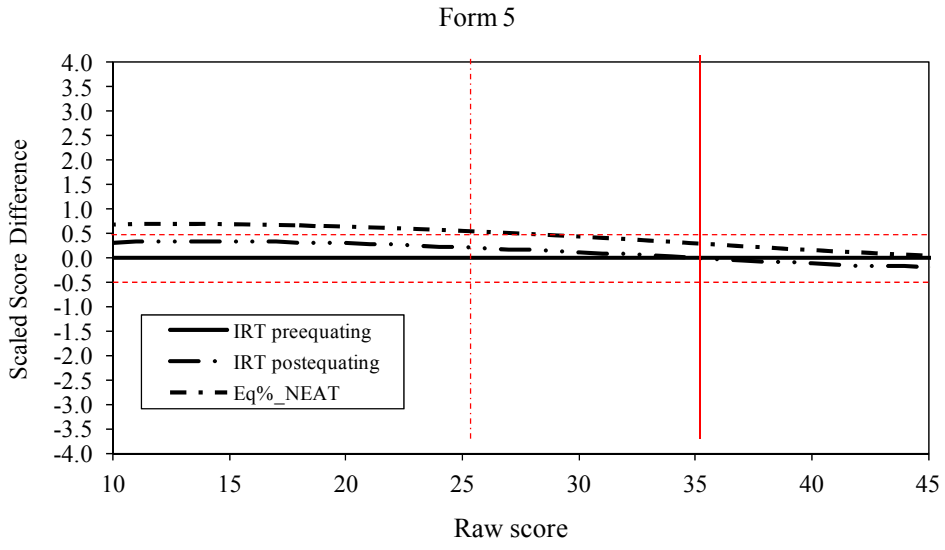

Figure 7. **Scaled score difference for ALG Form 5. Eq%_NEAT = chained-equipercentile equating; IRT = item response theory.**

IRT true score preequating yielded slightly overall lower passing rate (32.8%) than observed score equating (33.1%) and IRT true score postequating (33.5%). This is consistent with the descriptive statistics of the average scaled scores, that is, IRT true score preequating tends to yield the lowest average scaled scores (51.0) than the observed score equating method (51.7) and the IRT true score postequating (51.3).

Table 3 presents the three indices used to evaluate the equating results with IRT true score preequating results as the baseline. All three indices indicate that IRT true score postequating yields results closer to the IRT true score preequating methods, having smaller RMSD, MSD, and MAD in almost all forms.

Table 3

RMSD, MSD, and MAD for College Algebra (ALG)

	RMSD		MSD		MAD	
Form	Obs.	IRT post	Obs.	IRT post	Obs.	IRT post
1	1.091	0.731	0.986	0.708	0.986	0.708
2	0.376	0.455	0.317	0.429	0.320	0.429
3	1.299	0.423	1.226	0.390	1.226	0.390
4	0.695	0.291	0.665	0.260	0.667	0.260
5	0.483	0.211	0.435	0.126	0.435	0.184
6	0.231	0.135	0.195	-0.031	0.195	0.118
7	1.401	0.456	1.309	0.402	1.350	0.411
8	1.635	0.358	1.290	0.324	1.474	0.324

Note. IRT post = item response theory true score postequating; MAD = mean absolute difference; MSD = mean signed difference; Obs = observed score equating; RMSD = root mean square difference.

Conclusion and Discussion

For AIL, the reason for higher scaled scores from IRT true score preequating than from the two postequating methods can be partly attributed to how pretest items were tested, or, as we called it, *the scrolling effect*. When items are pretested, the items attached to the same stimulus are often embedded in a longer set, containing five to six items, than they do operationally, which may contain two to three items. Because the test was administered through a computer, examinees may have needed to scroll back to a few more pages to find the relevant information in the stimuli to answer the items and, thus, may have had more difficulty. As a result, the items might appear harder at the pretest stage than at the operational stage. However, for those

examinees whose scores were higher, their performance was less likely affected by the scrolling effect, as their scaled scores given by different equating methods were only slightly different. The differences in most forms should not be of concern, as they were within the bands allowed by the DTM.

The results for GOV are not as consistent as those for AIL. The scaled scores from IRT true score preequating are about the same as those from IRT true score postequating on three forms, but lower than those from IRT true score postequating on four forms. The examination of the items included in each form revealed that the former three forms had the same 31 items, and the latter four forms had a different group of 31 items. Interestingly, about half the items in the latter group of 31 items appeared to be slightly easier at the pretest stage than at the operational stage, and that might have contributed to the lower scaled scores in the IRT true score preequating method. We suspect that this result may be related to contextual/order effects.

For ALG, the reason for the lower scaled scores from IRT true score preequating than for the two postequating methods can be partly attributed to the speed of the test. Examinee responses from those who were not able to reach the pretest items were excluded from item calibration. As examinees of lower ability were usually those who were excluded, the calibration based on the data from higher ability examinees would make the items appear easier and scaled scores lower at the pretest stage.

Our results also indicate that, in general, the results from IRT true score postequating are closer to IRT true score preequating than those from observed score equating. We suspect that the difference between IRT true score equating and observed score equating may be due to the violation of IRT assumptions. In this testing program, tests do not always consist of stand-alone independent items. Rather, item bundles consisting of an aggregation of items on a single stimulus are used in most of the test forms. When item bundles are used in the test, but the unidimensional IRT model is still applied to modeling response data, the IRT assumption of local independence may be violated, causing item parameter estimates to be problematic. As a result, IRT equating results are likely to be adversely affected (Chen & Thissen, 1997; Yen, 1984).

In actual practice, the selection of an equating design is an important part of the test development plan even before the tests are administered operationally. This excludes the possibility of conducting equating by choosing an equating design and an equating method that

can best fit the data structure, as we did in this study. In this respect, the way this study was conducted can be regarded as a type of post-hoc check aiming to provide insights into the degree of consistency in results from pre- and postequating. These insights can provide very useful information with regard to how well the preequating works in a testing program in comparison to postequating methods, as well as pointing to causes of the differences. We highly recommend the activities that were conducted in this study as an important quality check step for any testing program that employs IRT true score preequating. Meanwhile, due to the use of operational data, this study cannot provide a clear picture of how different factors, as well as the levels of differences in each factor, may affect the differences in pre- and postequating results. In this regard, a simulation study is greatly needed to study factors that could not be studied in this research.

References

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

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Domaleski, C. (2006). *Exploring the efficacy of preequating a large scale criterion-referenced assessment with respect to measurement equivalence* (Unpublished doctoral dissertation). Georgia State University, Atlanta.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT® and PSAT/NMSQT®. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Deryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10, pp. 1–32). Princeton, NJ: ETS.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program® exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS Research Report No. RR-03-27, pp. 19–36). Princeton, NJ: ETS.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT® verbal and mathematical sections* (ETS Research Report No. RR-85-10). Princeton, NJ: ETS.

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT true-score preequating* (ETS Research Report No. RR-86-14). Princeton, NJ: ETS.

Kirkpatrick, R., & Way, W. D. (2008, April). *Field testing and equating designs for state educational assessments*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practice.* New York, NY: 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.

Livingston, S. A. (1985, April). *Large-sample preequating: How accurate?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT preequating* (ETS Research Report No. RR-86-49). Princeton, NJ: ETS.

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

Tong, Y., Wu, S.-S., & Xu, M. (2008, April). *A comparison of preequating and postequating using large-scale assessment data*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.