RESEARCH REPORT 2016-4

CollegeBoard

Evaluating Equity at the Local Level Using Bootstrap Tests

By YoungKoung Kim and Lawrence T. DeCarlo

YoungKoung Kim is a psychometrician in the Research Department of the College Board.

Lawrence T. DeCarlo is a professor of psychology and education in the Department of Human Development, Teachers College, Columbia University

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators, and schools. For further information, visit www.collegeboard.org.

© 2016 The College Board. College Board, Advanced Placement Program, AP, SAT, and the acorn logo are registered trademarks of the College Board. PSAT/NMSQT is a registered trademark of the College Board and National Merit Scholarship Corporation. All other products and services may be trademarks of their respective owners. Visit the College Board on the Web: www.collegeboard.org.

RESEARCH

For more information on College Board research and data, visit research.collegeboard.org.

Contents

Tables

Figures

Introduction

Because of concerns about test security, different test forms are typically used across different testing occasions. As a result, equating is necessary in order to get scores from the different test forms that can be used interchangeably. In order to assure the quality of equating, multiple equating methods are often examined. Various equity properties have been used to assess the adequacy of different equating methods. For example, following the "weak" definition of equity (Morris, 1982), first-order equity refers to two alternate forms having similar expected scores given true examinee ability. Second-order equity refers to two alternate forms having a similar error of measurement given true examinee ability. Thus, first- and second-order equity properties examine the first and the second central moments of the conditional distributions of test scores at given levels of examinee ability.

The discrepancy indices D_1 and D_2 (defined in more detail below) are often used as evaluation criteria for satisfying equity properties (Kim, Brennan, & Kolen, 2005; Tong & Kolen, 2005; Lee, Lee, & Brennan, 2010). These discrepancy indices evaluate equity at the global level as a form of weighted average difference in the expected scale scores or error variance of measurement between alternate forms over all levels of examinee ability. Thus, these indices provide an overall summary of whether equity properties are satisfied between two alternate forms over the entire range of scale scores.

If there is a specific range of scale scores that requires specific attention (for example, cut scores for licensure tests and scholarship competitions), then it would be advisable to evaluate equity at the local rather than the global level. At the local level, however, there are no formal statistical tests or evaluation criteria available to assess equity, apart from examining graphs. Furthermore, it is challenging to conduct formal statistical tests of the differences in the conditional mean and variance between two test forms because the statistical tests would require a strong assumption about the conditional distribution of the test scores at each examinee ability level. Therefore, the current study proposes a method that evaluates equity properties using the bootstrap technique, which allows for a statistical test of equity at the local level without making any distributional assumptions.

The bootstrap tests for the equity properties are developed by following the bootstrap procedure discussed in Boos and Brownie (1989). The current study demonstrates the bootstrap tests using large-scale assessment data in which the scores are used to determine eligibility for scholarship for students. The equity properties for the assessment, in particular the cut scores in which the scholarships are determined, are examined within an IRT framework (Kolen, Zeng, & Hanson, 1996). In terms of equating design, the current study focuses on the common item nonequivalent groups equating design. Four equating methods — IRT true score, IRT observed score, frequency estimation (FE), and chained equipercentile (CE) method — are compared.

Evaluating Equity Properties and Evaluation Criteria

If equating is appropriate, the distribution of scale scores should be the same on the old form and the new form after equating. Thus, the adequacy of equating is often evaluated by assessing two equity properties — first-order equity and second-order equity. The equity properties hold if examinees at a given true score have the same expected scale score

(first-order equity) and the same conditional error of measurement (second-order equity) on the two forms (Tong & Kolen, 2005).

For tests with dichotomously scored items, several procedures have been developed to examine equity properties (Kolen, Hanson, & Brennan, 1992; Kolen et al., 1996). Kolen et al. (1996) presented a procedure for estimating the expected values and conditional standard errors of measurement of scale scores in an IRT framework. According to Kolen et al., for a test with K items, the first-order equity refers to the true expected scale score $\xi(\theta)$ at given examinee ability θ, which can be written as

$$\xi(\theta) = E[s(X) \mid \theta] = \sum_{i=0}^{K} s(i) P(X = i \mid \theta),$$

in which $s(X)$ is the raw to scale score transformation function, and $P(X = i \mid \theta)$ is the probability of the raw score random variable X being equal to i ($i = 0, 1, \ldots K$), which can be calculated using Lord and Wingersky's recursive formula (1984). Regarding second-order equity, the conditional standard error of measurement (CSEM) of a scale score given θ can be written as

$$\sigma(s(X) \mid \theta) = \sqrt{E\{[s(X) - \xi(\theta)]^2 \mid \theta\}} \sqrt{\sum_{i=0}^{K} [s(i) - \xi(\theta)]^2 P(X = i \mid \theta)}.$$

After estimating expected scale scores and CSEMs, previous research used the overall discrepancy indices — D_1 and D_2 — as evaluation criteria to demonstrate whether the equity properties had been preserved (Tong & Kolen, 2005; Lee et al., 2010). D_1 refers to the expected scale scores, and D_2 refers to the CSEMs. By following Lee et al. (2010) for two forms — Form W and Form S — D_1 and D_2 can be written as follows:

$$D_1 = \frac{\sum_i w_i \mid E(SC_W \mid \theta_i) - E(SC_S \mid \theta_i) \mid}{\sum_i w_i}$$

$$D_2 = \sqrt{\frac{\sum_i w_i \mid (EV_W \mid \theta_i) - (EV_S \mid \theta_i) \mid}{\sum_i w_i}}$$

in which $E(SC_w|\theta_i)$ and $E(SC_s|\theta_i)$ are the expected scale scores from the two forms, Form W and Form S, at quadrature point θ_i, $EV_w|\theta_i$ and $EV_s|\theta_i$ are the conditional error variance of measurements for the two forms at quadrature point θ_i and w_i is the weight of θ_i. Smaller values of D_1 and D_2 imply small discrepancies between two forms in terms of the first-order equity and the second-order equity, and thus imply that the equity properties are preserved.

Although D_1 and D_2 provide information on whether the equity properties are preserved, the discrepancy indices only summarize the information at the global level, that is, over the entire range of the examinee ability distribution. If a particular level of examinee ability is of interest, these overall summary indices do not provide this information. Instead of a statistical test, graphical inspection at the examinee ability level of interest may be used. In practice, ensuring the preservation of equity properties at particular examinee ability levels is often of great importance for test fairness when the ability levels are used for cut scores of, among other things, a licensure test, scholarship competition, or college admission. In those cases, a formal statistical test for assessing these equity properties is beneficial.

Bootstrap Tests for Equity Properties

Boos and Brownie (1989) suggested bootstrap test procedures introduced by Efron (1979) to test for the homogeneity of variance without making any assumptions about the sample distribution. For example, the commonly used F test of equality variance, S^2_1/S^2_2, requires a normal distribution assumption. If the distribution is unknown, however, the test may not be appropriate. Thus, Boos and Brownie proposed a bootstrap technique that allows one to use normal theory test statistics, such as the F test for equal variance between two samples, without making a normality assumption.

The bootstrap technique by Boos and Brownie is applicable to testing for equity properties, in particular, testing for second-order equity, which is related to the CSEMs from two test forms. To test whether the second-order equity holds, the variance of the conditional scale score distribution at a given level of examinee ability in one test form needs to be compared with the variance in the other test form. Since the parametric form of the conditional scale score distribution given examinee ability is not necessarily known, the bootstrap procedure can be applied to test the equality of CSEMs from the two test forms.

Based on the bootstrap procedure discussed in Boos and Brownie, the proposed bootstrap tests for first-order equity and second-order equity are as follows:

- Draw k independent samples from the original sample with replacement.
- Compute the difference test statistics T_1 and T_2 for each bootstrap sample, and obtain the distributions of T_1 for first-order equity and T_2 for second-order equity:
  - $T_1 i = E(SC_j|\theta) - E(SC_2|\theta),\ i = 1, 2, \ldots k$
  - $T_2 i = (EV_j|\theta) / (EV_2|\theta),\ i = 1, 2, \ldots k$

  in which $E(SC_j|\theta)$ is the expected mean given examinee ability θ for form j and $(EV_j|\theta)$ is the error variance of measurement at θ for form j.
- Construct a 95% interval for T_1 and T_2 using the k samples.
- Examine whether the interval includes zero for the first-order equity and one for the second-order equity.

Method

Data

The current study demonstrates the bootstrap tests for the equity properties using a random sample from a pre-2015 PSAT/NMSQT® test administration. The PSAT/NMSQT assessment is a norm-referenced test designed primarily for 10th- and 11th-grade students. The pre-2015 PSAT/NMSQT includes three test areas: critical reading, mathematics, and writing. The primary intended uses are as a low-stakes assessment in preparation for taking the SAT® and as a high-stakes assessment for 11th-grade students to determine eligibility to participate in the National Merit Scholarship competition. The PSAT/NMSQT is a co-sponsored assessment between the College Board and the National Merit Corporation. The PSAT/NMSQT is administered in the fall, on a Wednesday and following Saturday of the same week, with separate forms.

The pre-2015 PSAT/NMSQT inherits many of its content and psychometric properties from the SAT. The reported PSAT/NMSQT score scale ranged from 20 to 80 in increments of 1, for a total of 61 scale score points. The PSAT/NMSQT score scale was primarily maintained through the maintenance of the SAT score scale. The parent pre-March-2016 SAT forms are equated to previously administered SAT forms. Once the pre-2015 PSAT/NMSQT forms are administered, they were then equated back to their parent pre-March-2016 SAT forms using the common item nonequivalent groups equating design. At least 60%–65% of the total items in the pre-2015 PSAT/NMSQT are used as common items. See Cook, Dunbar, & Eignor (1981) for details on the equating design of the pre-2015 PSAT/NMSQT forms.

Analyses

For each pre-2015 PSAT/NMSQT test form (Form W and S), a random sample with N = 10,000 was obtained. For its SAT parent forms, a random sample with N = 5,000 was used. To conduct the bootstrap tests, the following steps were used:

- Step 1: For each form, a bootstrap sample was drawn from the original sample with replacement.
- Step 2: The items in the bootstrap sample were calibrated using flexMIRT version 2.0 (Cai, 2013). The 3 parameter logistic IRT model was used for item calibration.
- Step 3: Once item calibration was completed, items from the two PSAT/NMSQT forms were placed on the SAT scale using the Haebara method (Haebara, 1980).
- Step 4: Equating was conducted using four methods — IRT true score, IRT observed score, frequency estimation (FE), and chained equipercentile (CE) method.
- Step 5: Based on the equating results, the expected rounded scale scores and conditional standard errors of measurement were computed to examine the first-order equity and the second-order equity. The difference test statistics, T_1 and T_2, were computed.
- Step 6: Step 1 to 5 were repeated 1,000 times.

Results

The descriptive statistics of the raw scores from the random samples of the pre-2015 PSAT/NMSQT (N = 10,000) and pre-March-2016 SAT (N = 5,000) forms are given in Table 1. In addition, the effect size[1] between the PSAT/NMSQT and SAT forms is presented. While the main test populations of the PSAT/NMSQT are based on 10th- and 11th-graders, the SAT primary test populations are from 11th- and 12th-graders. Thus, the examinee ability of the SAT takers tends to be higher than that of the PSAT/NMSQT test-takers. Due to this difference in test-taker populations between the PSAT/NMSQT and SAT, the observed effect sizes between the PSAT/NMSQT and SAT were quite large.

1. The effect size was computed using the following:

$$\text{Effect Size} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(s_1^2 + s_2^2\right)/2}}$$

Table 1.

Summary Statistics for Raw Scores

Test Section	PSAT/NMSQT					SAT					Effect Size
	Form W					SAT 1					
	N	*Mean*	*SD*	*Min*	*Max*	*N*	*Mean*	*SD*	*Min*	*Max*	
Critical Reading	10,000	22.76	8.63	1	48	5,000	34.82	12.45	6	67	-1.13
Writing	10,000	18.64	7.88	1	39	5,000	27.30	9.01	5	49	-1.02
Math	10,000	17.12	8.11	1	38	5,000	28.39	11.26	2	54	-1.15
	Form S					SAT 2					
	N	*Mean*	*SD*	*Min*	*Max*	*N*	*Mean*	*SD*	*Min*	*Max*	
Critical Reading	10,000	28.36	8.16	5	48	5,000	35.54	12.58	5	67	-0.68
Writing	10,000	23.44	7.11	3	39	5,000	27.30	9.23	3	49	-0.47
Math	10,000	21.73	6.96	2	38	5,000	29.24	11.11	3	54	-0.81

The two PSAT/NMSQT forms were equated to their SAT parent forms, which had been already equated, using a common-item non-equivalent groups design; that is, the PSAT/NMSQT Form W was equated to SAT Form 1, and the PSAT/NMSQT Form S was equated to SAT Form 2. Table 2 shows the discrepancy indices — D_1 and D_2 — for the results of the rounded scale score equating using four equating methods. The IRT equating (true score and observed score equating) methods seem better than the FE and CE methods because the IRT methods had smaller D_1 values than those of either the FE or CE method for all three sections. On the other hand, the FE and CE methods had smaller D_2 values than the IRT methods. Thus, the two IRT methods seemed to preserve the first-order equity better than the FE and CE methods while the FE and CE methods seemed to preserve the second-order equity better than the IRT methods. The second-order equity seemed to be slightly better preserved for the Critical Reading and Writing than for Math.

Table 2.

Discrepancy Indices for Rounded Scale Score Equating

Discrepancy Index	D_1			D_2		
Section	CR	M	W	CR	M	W
IRT True	0.31	0.23	0.23	1.01	1.43	0.87
IRT Obs	0.32	0.11	0.18	0.96	1.27	0.86
FE	0.41	0.60	0.27	0.80	1.26	0.90
CE	0.37	0.54	0.24	0.83	1.25	0.77

The plots of expected scale scores for both PSAT/NMSQT forms as well as the difference in the expected scale scores between the two forms are shown in Figures 1 to 4 for Critical Reading, Figures 13 to 16 for Math, and Figures 25 to 28 for Writing. Overall, the differences in expected scores between the two forms were between –1 and +1 across all examinee ability levels for all equating methods except for the FE and CE methods in Writing. This small overall discrepancy of the expected scale scores between the two forms indicates that the first-order equity was well preserved. Although the overall difference was small, where the most differences occurred in the range of examinee ability varied across equating methods.

Figure 1.

First order equity using IRT true score equating for Critical Reading.

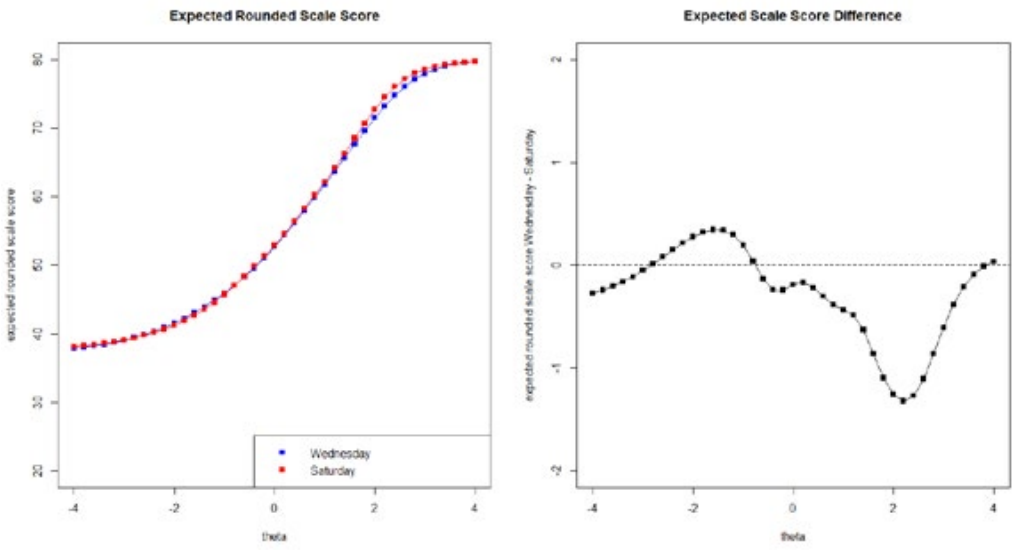


Figure 2.

First order equity using IRT observed score equating for Critical Reading.

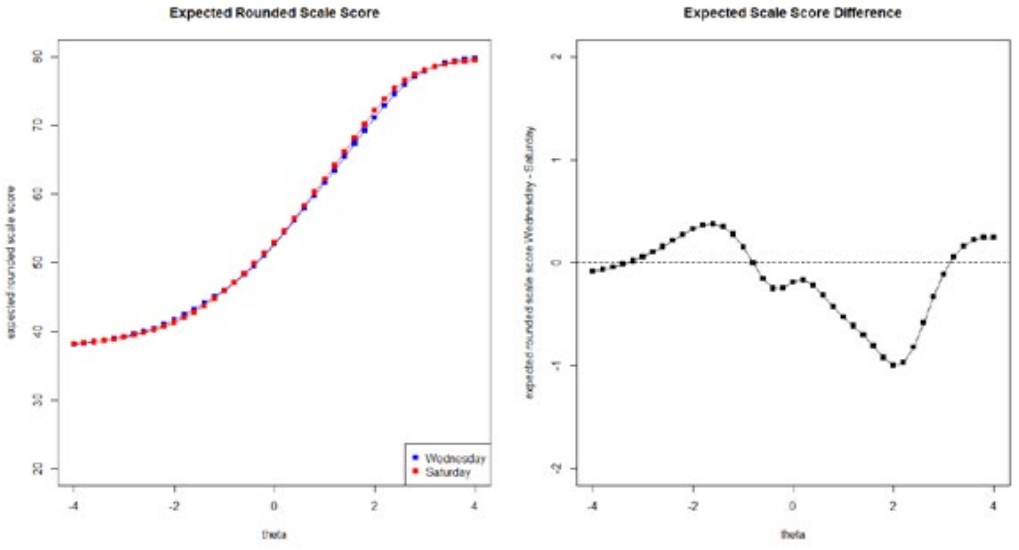


Figure 3.

First order equity using frequency estimation equating for Critical Reading.

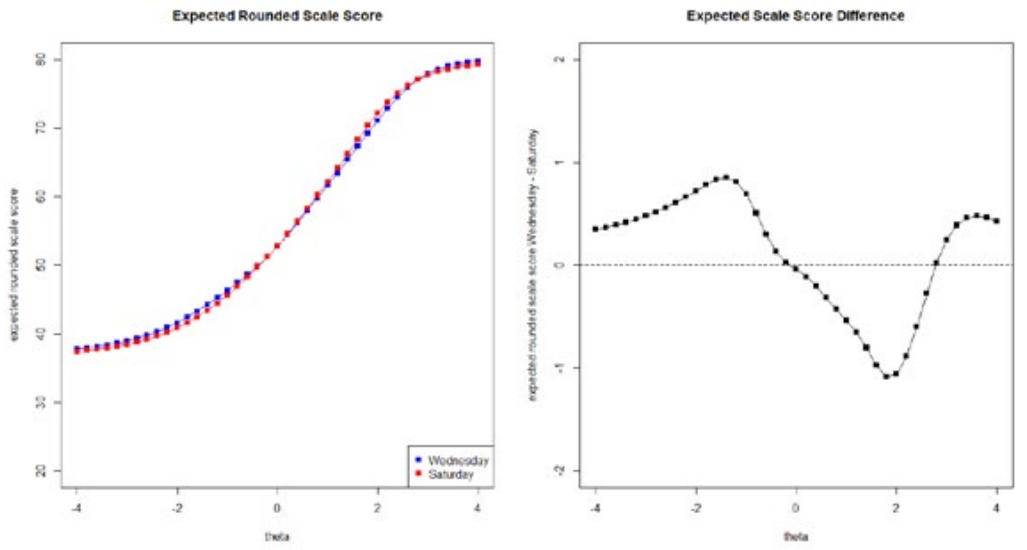
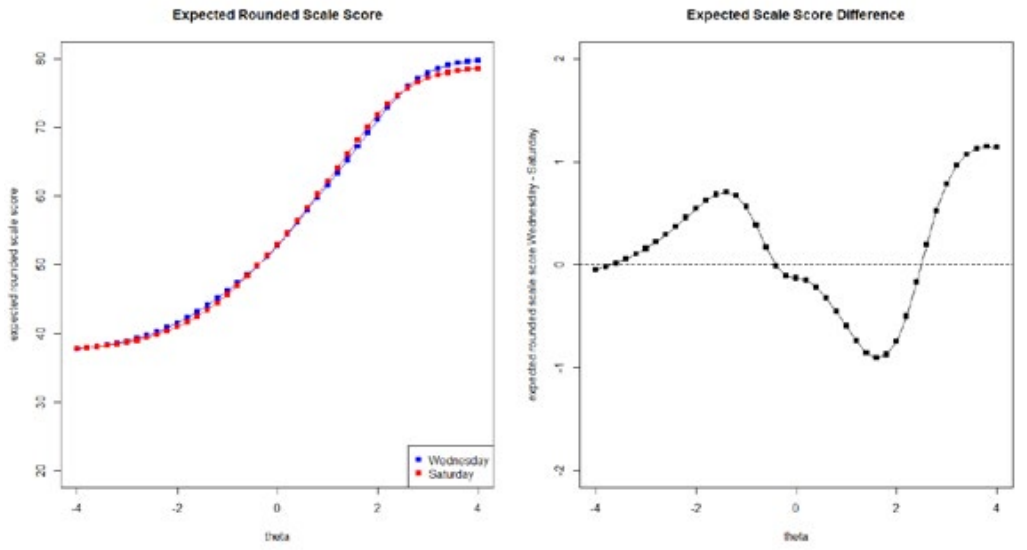


Figure 4.

First order equity using chained equipercentile equating for Critical Reading.



The plots of CSEMs for the two forms and the ratio of the CESMs between two forms are presented in Figures 5–8 for Critical Reading, Figures 17–20 for Math, and Figures 29–32 for Writing. As with the expected scale score difference, the difference in CSEMs between the two forms was small — the ratio ranges between 0.5 and 1.5 for all equating methods. The ratio tended to be larger at the higher level of the examinee ability.

Figure 5.

Second order equity using IRT true score equating for Critical Reading.

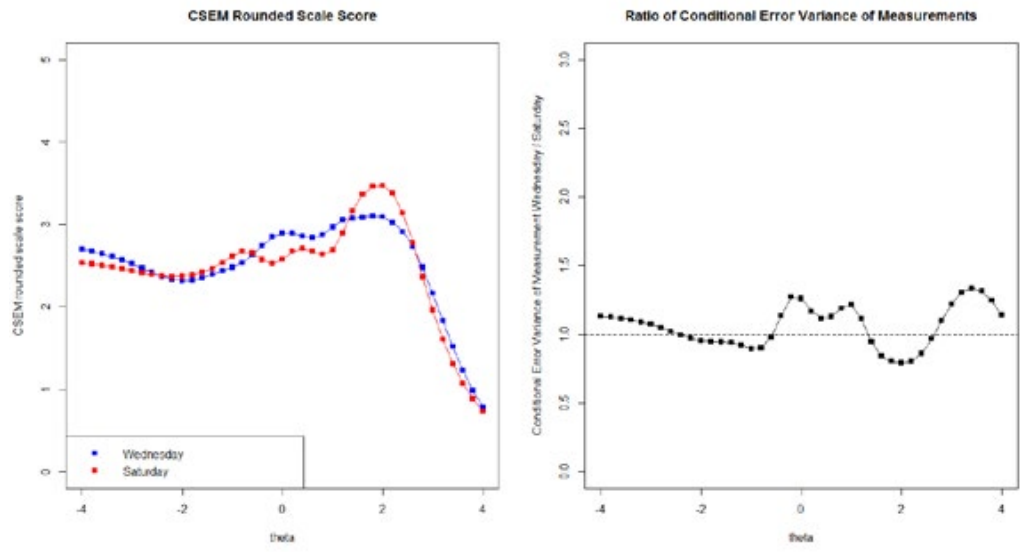


Figure 6.

Second order equity using IRT observed score equating for Critical Reading.

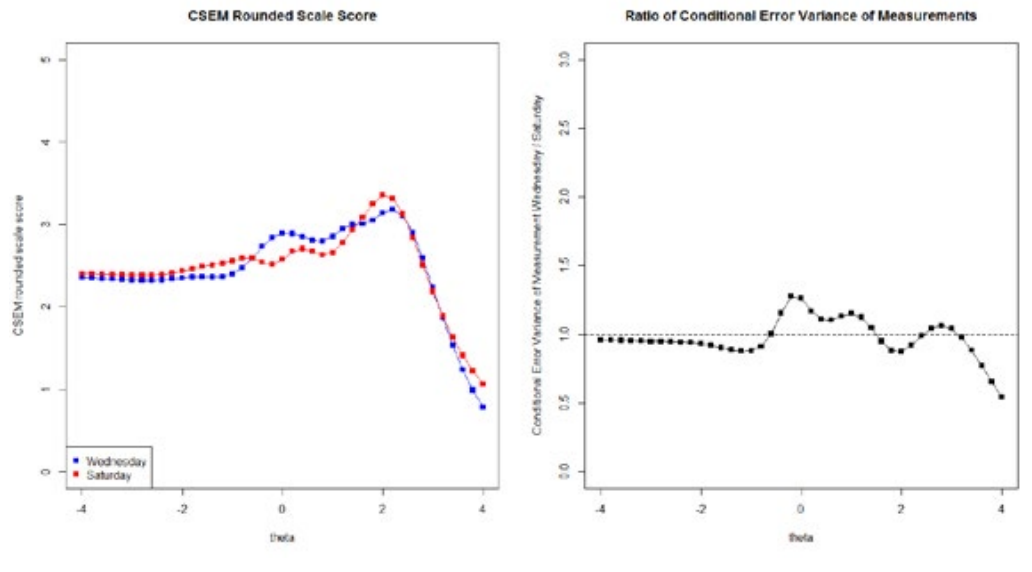


Figure 7.

Second order equity using frequency estimation equating for Critical Reading.

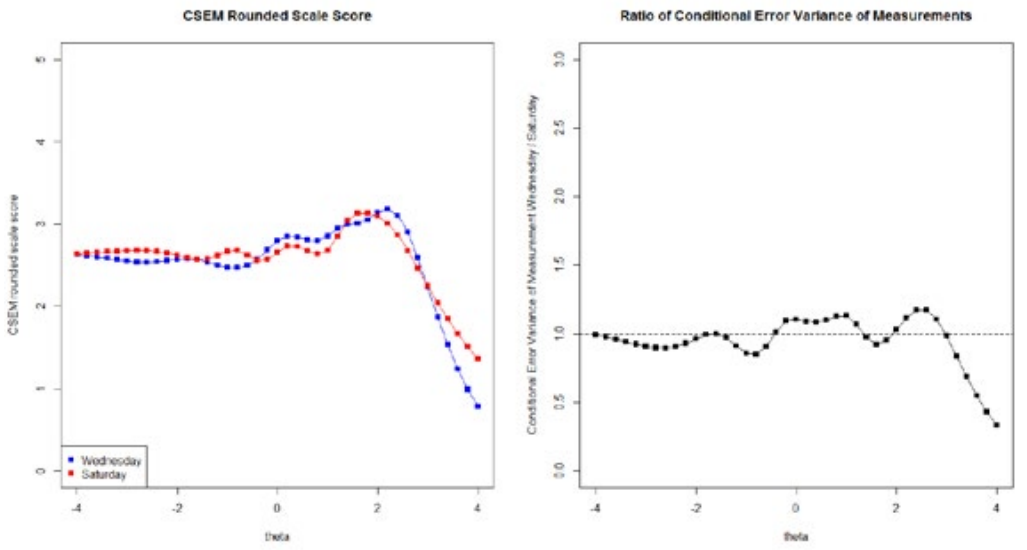
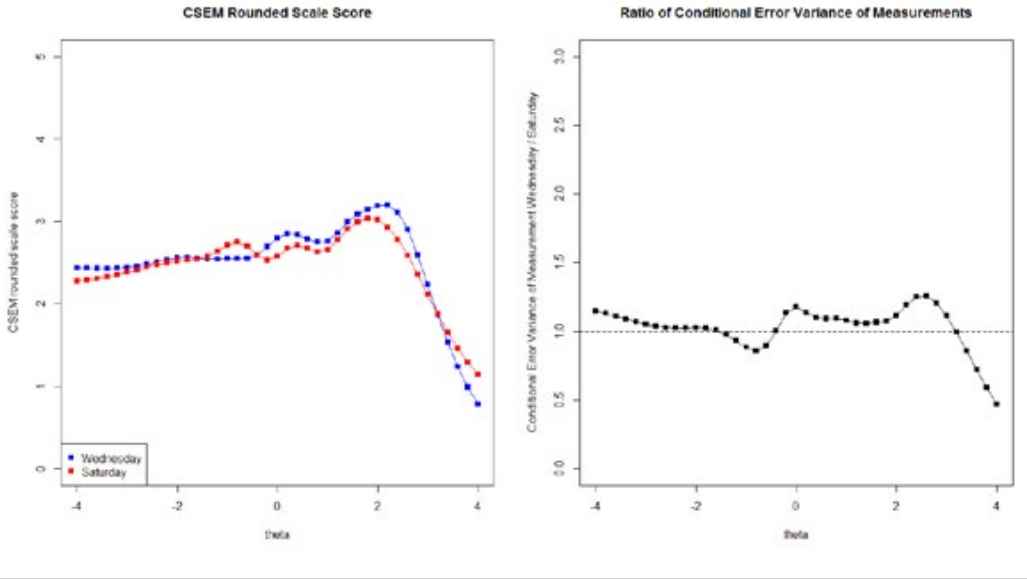


Figure 8.

Second order equity using chained equipercentile equating for Critical Reading.



After examining the plots of expected scale scores and the ratio of the CSEM, a natural question to ask is "Are these differences significant?" The bootstrap test can answer this question. Tables 3 and 4 summarize the results of the bootstrap tests for equity properties. Figures 9–12 for Critical Reading, Figures 21–24 for Math, and Figures 33–36 for Writing show

the 95% confidence interval for the first-order equity and the second-order equity bootstrap tests. Using 95% confidence interval band plots, the local examinee ability levels that do not contain zero for the first-order equity, and one for the second-order equity can be easily detected for each equating method.

Figure 9.

Bootstrap test for equity for IRT true score equating for Critical Reading.

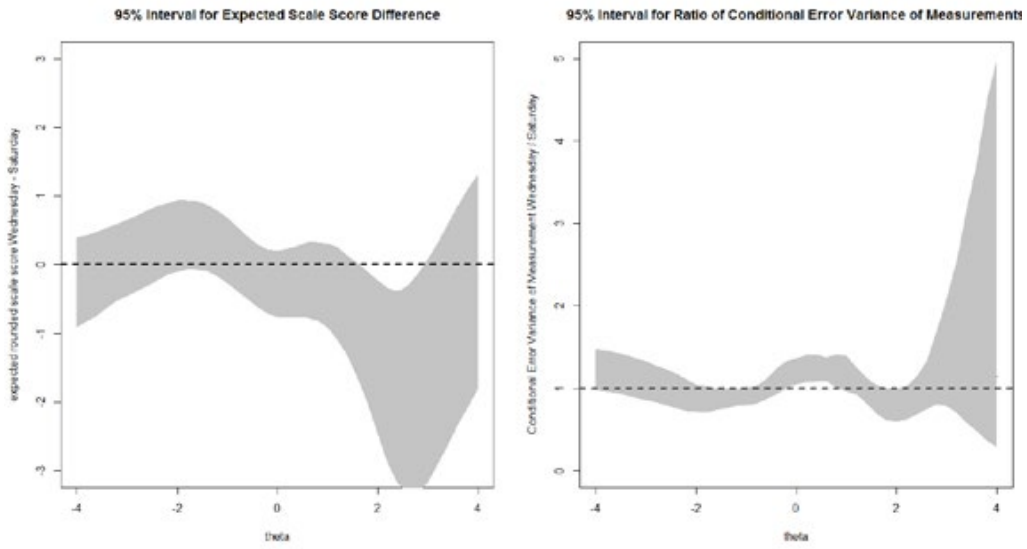


Figure 10.

Bootstrap test for equity for IRT observed score equating for Critical Reading.

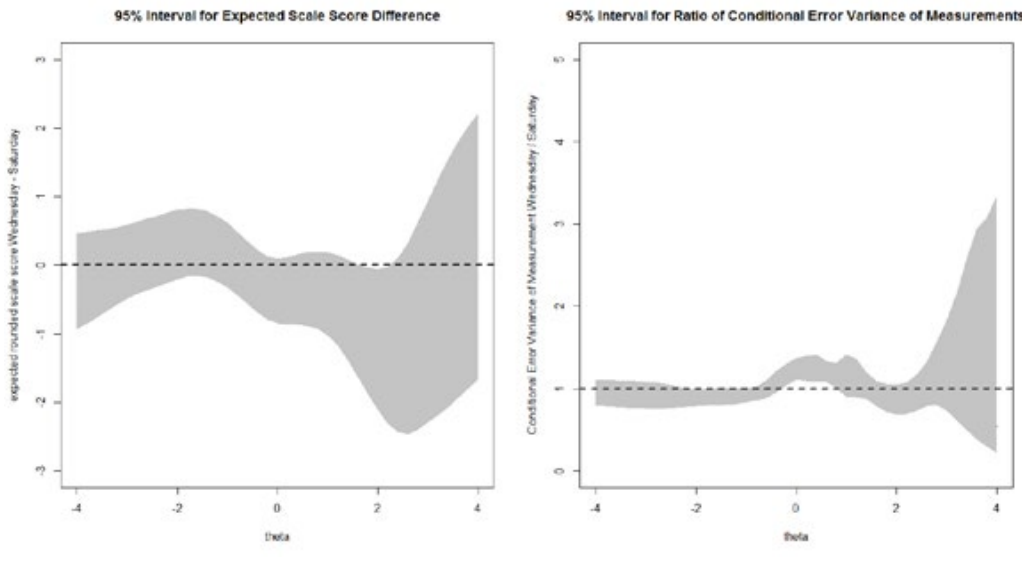


Figure 11.

Bootstrap test for equity for frequency estimation equating for Critical Reading.

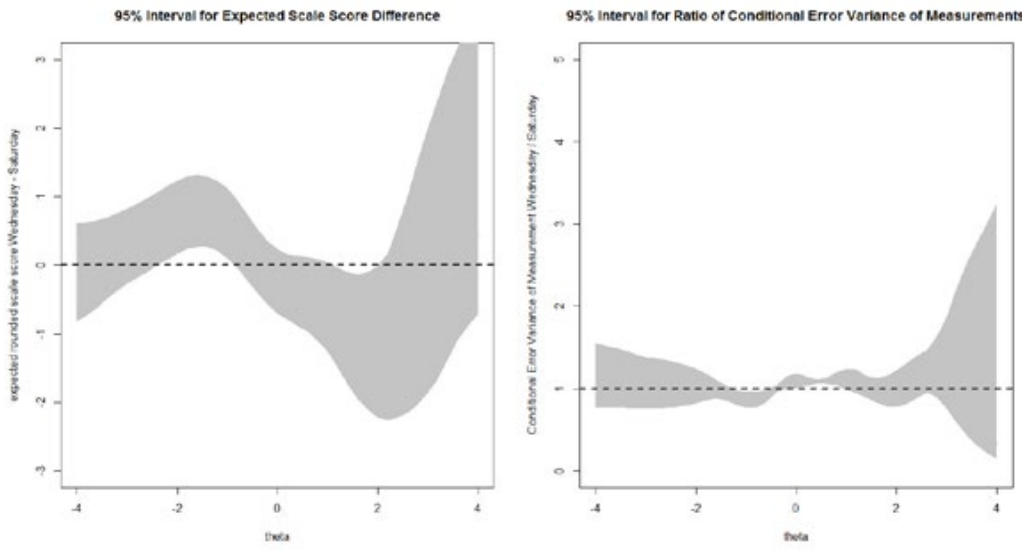


Figure 12.

Bootstrap test for equity for chained equipercentile equating for Critical Reading.

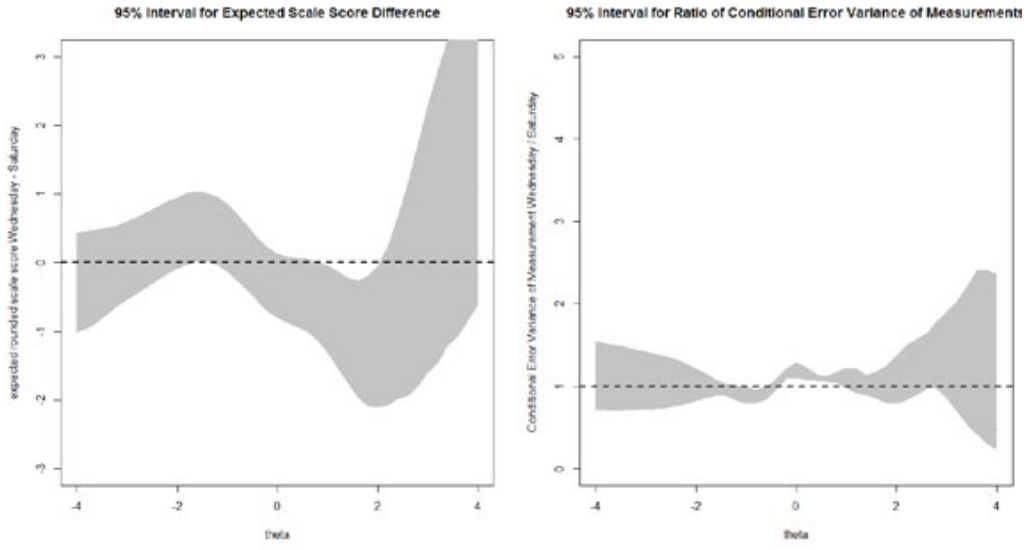


Figure 13.

First order equity using IRT true score equating for Math.

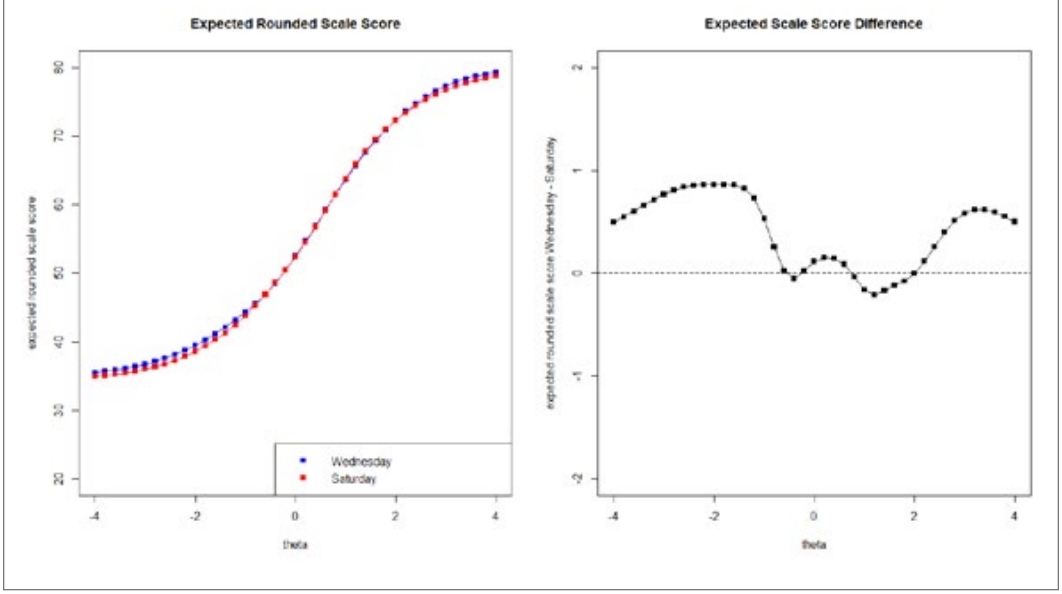


Figure 14.

First order equity using IRT observed score equating for Math.

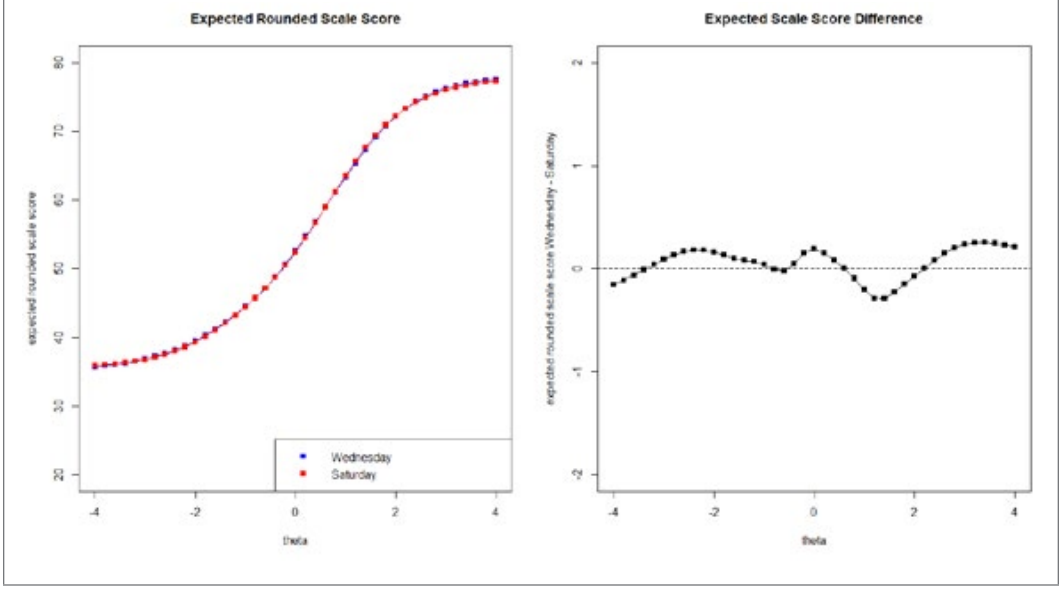


Figure 15.

First order equity using frequency estimation equating for Math.

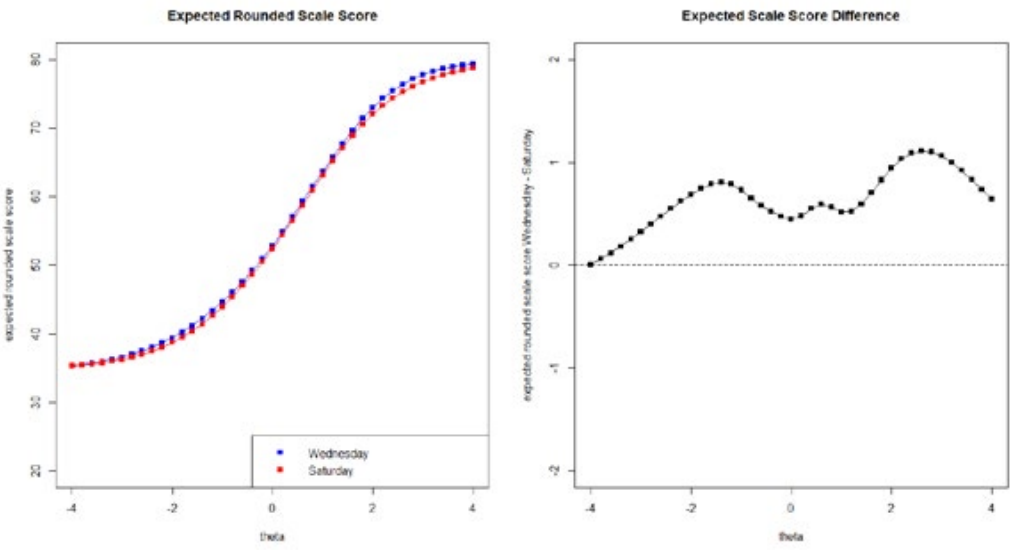


Figure 16.

First order equity using chained equipercentile equating for Math.

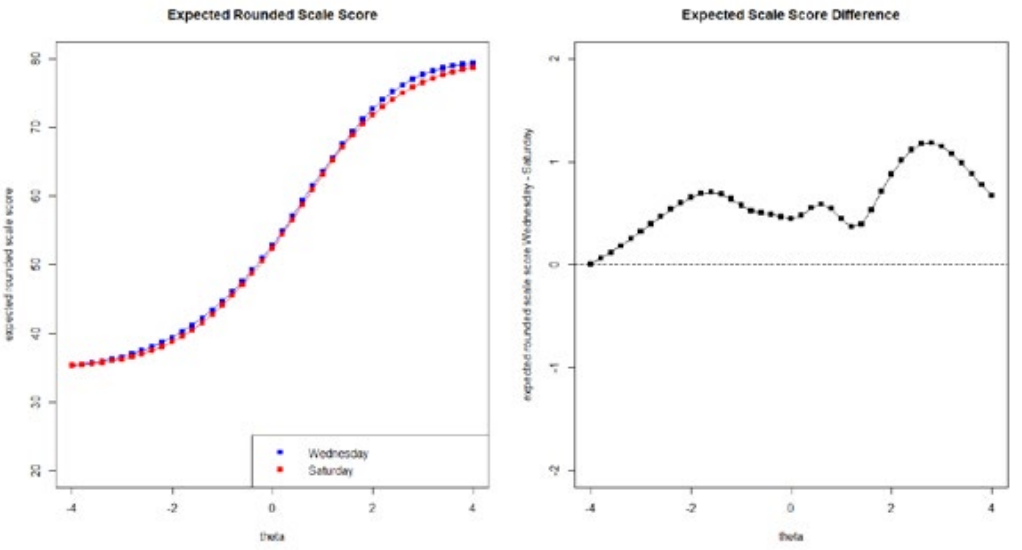


Figure 17.

Second order equity using IRT true score equating for Math.

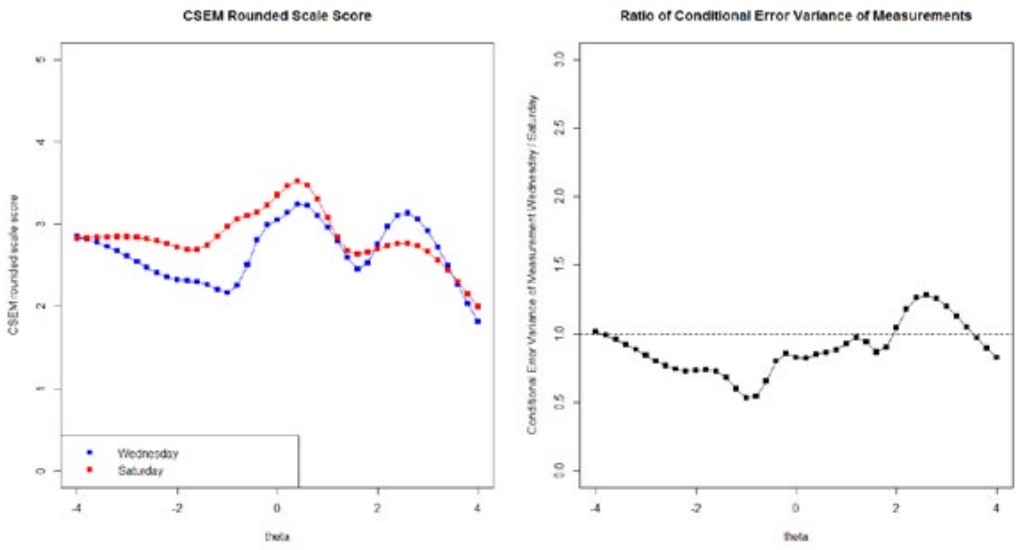


Figure 18.

Second order equity using IRT observed score equating for Math.

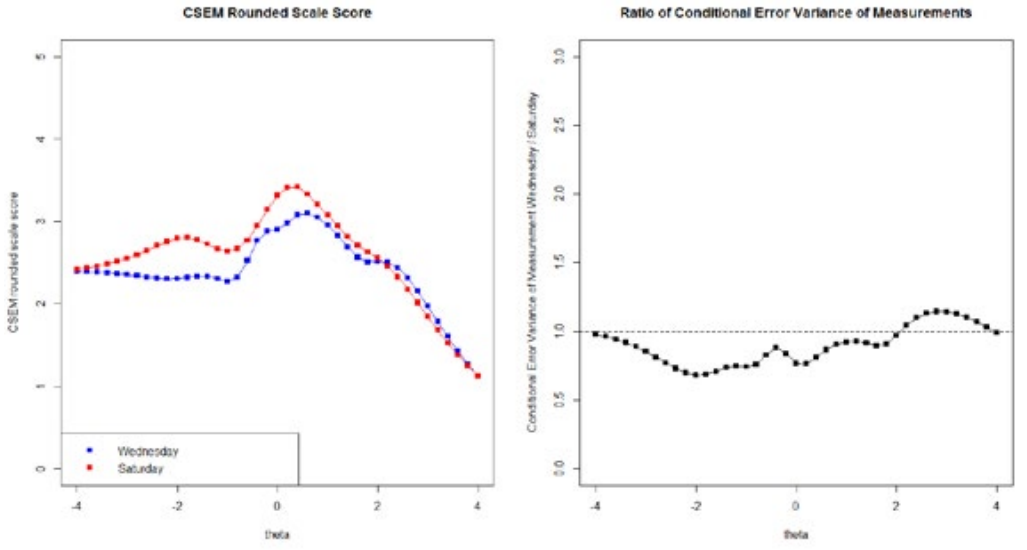


Figure 19.

Second order equity using frequency estimation equating for Math.

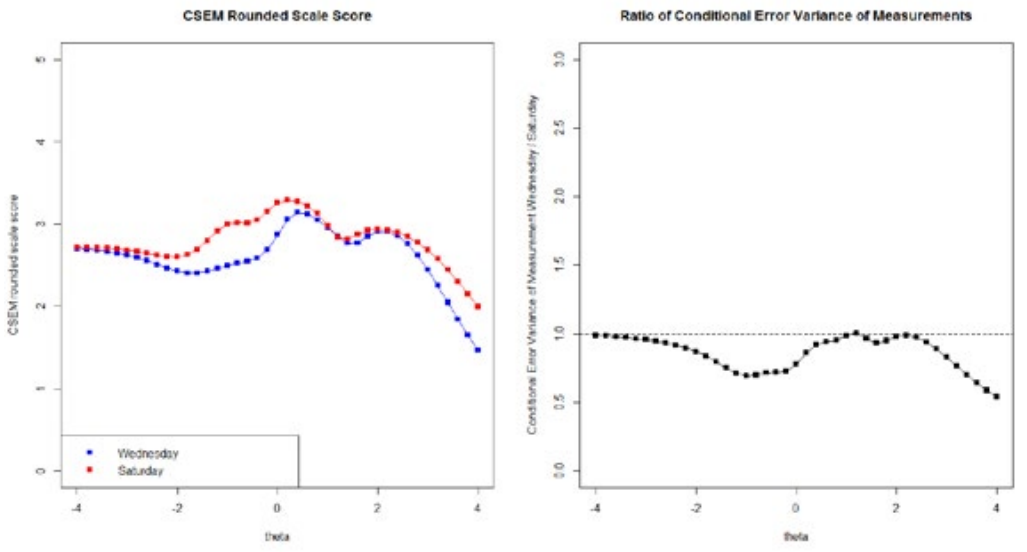


Figure 20.

Second order equity using chained equipercentile equating for Math.

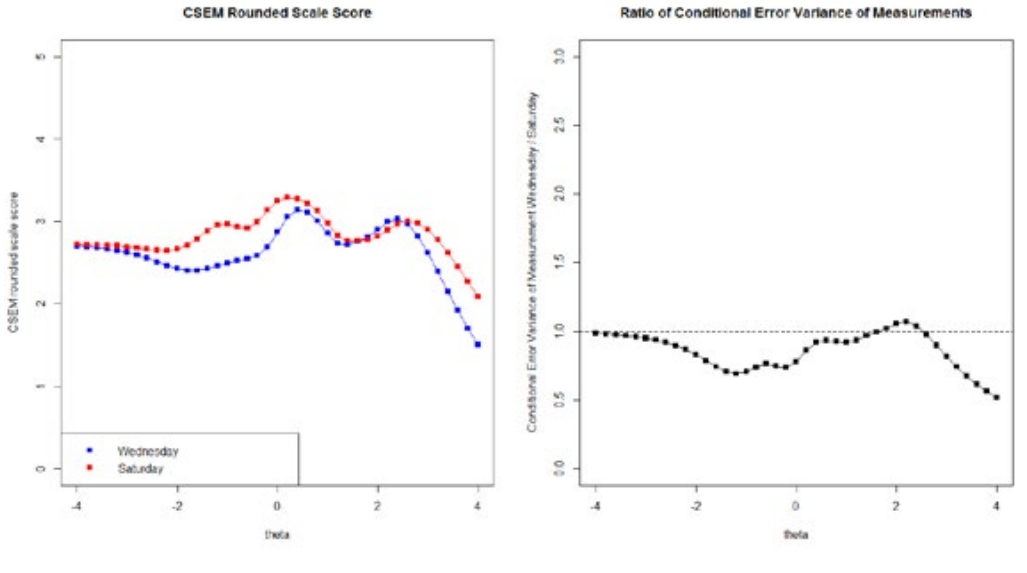


Figure 21.

Bootstrap test for equity for IRT true score equating for Math.

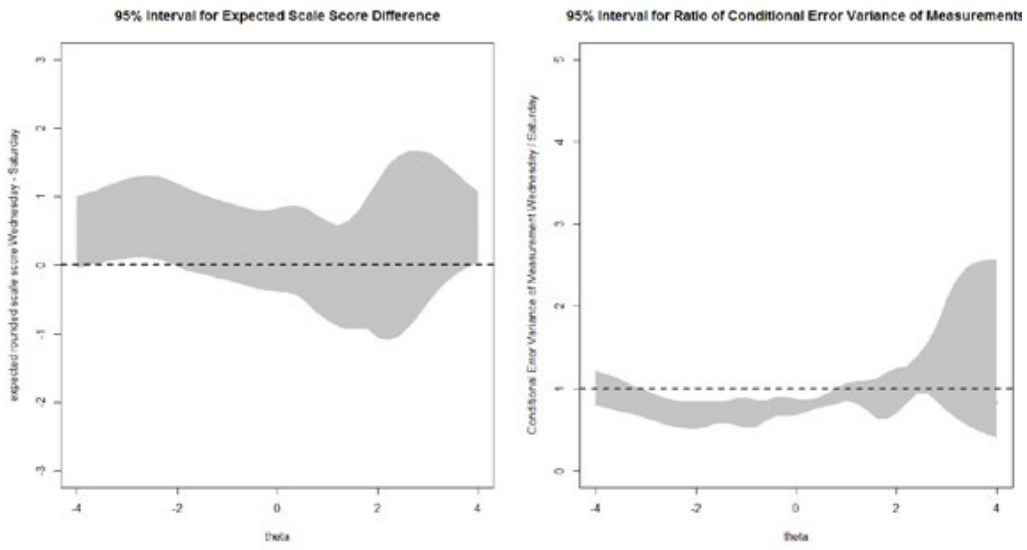


Figure 22.

Bootstrap test for equity for IRT observed score equating for Math.

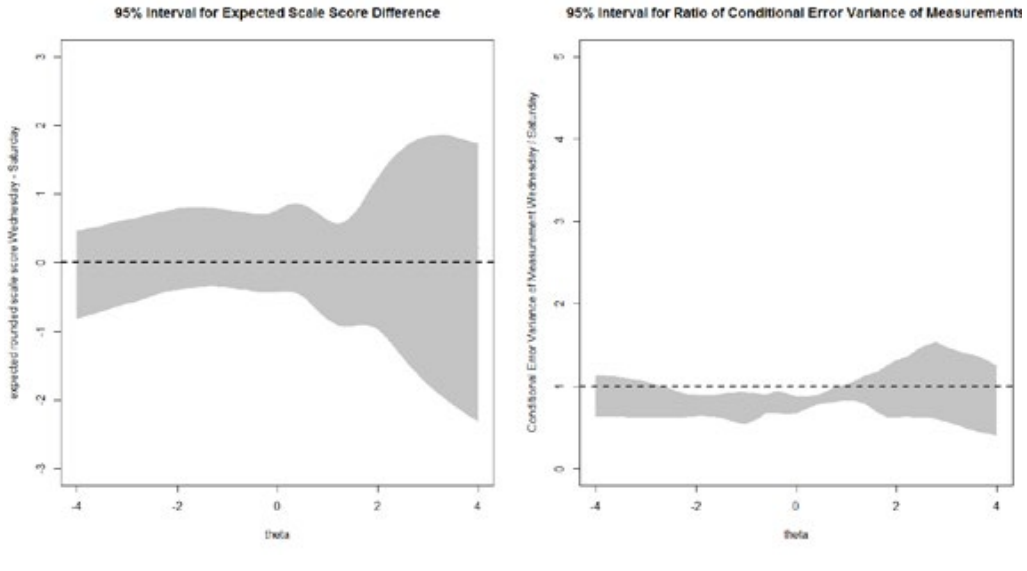


Figure 23.

Bootstrap test for equity for frequency estimation equating for Math.

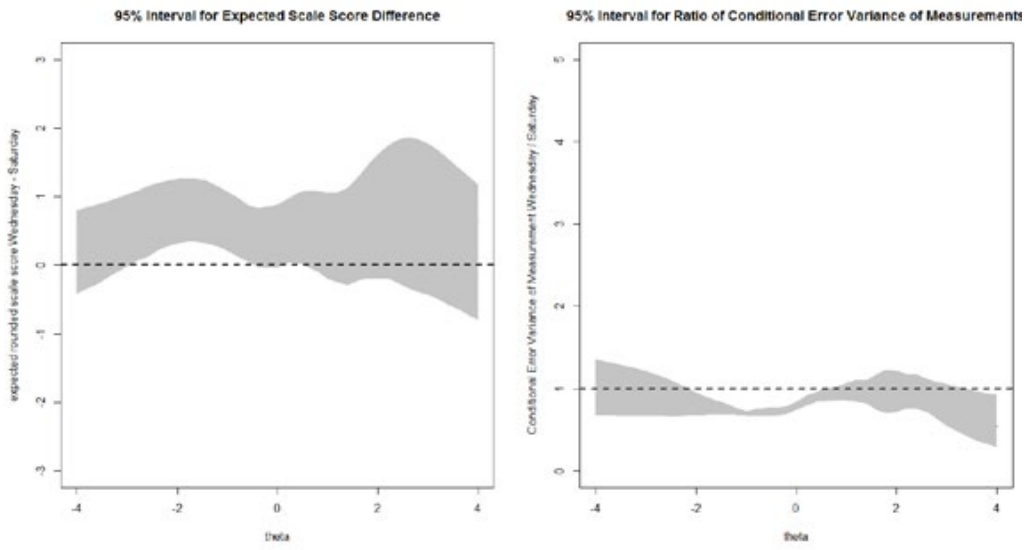


Figure 24.

Bootstrap test for equity for chained equipercentile equating for Math.

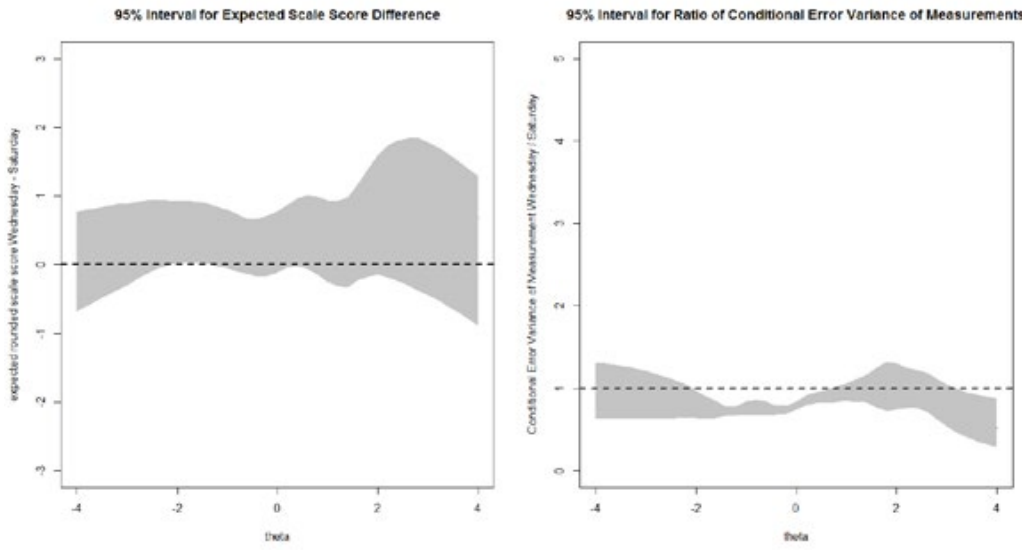


Figure 25.

First order equity using IRT true score equating for Writing.

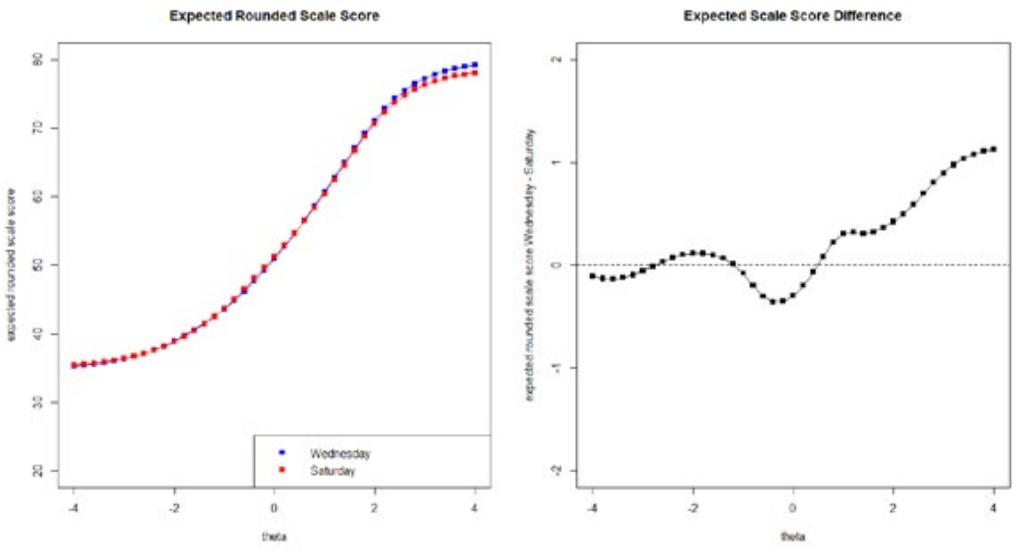


Figure 26.

First order equity using IRT observed score equating for Writing.

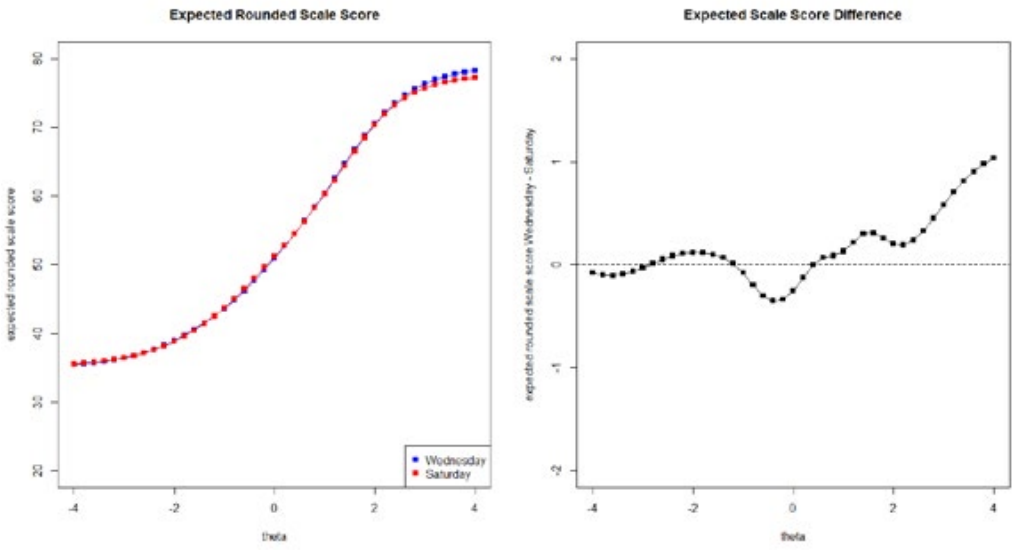


Figure 27.

First order equity using frequency estimation equating for Writing.

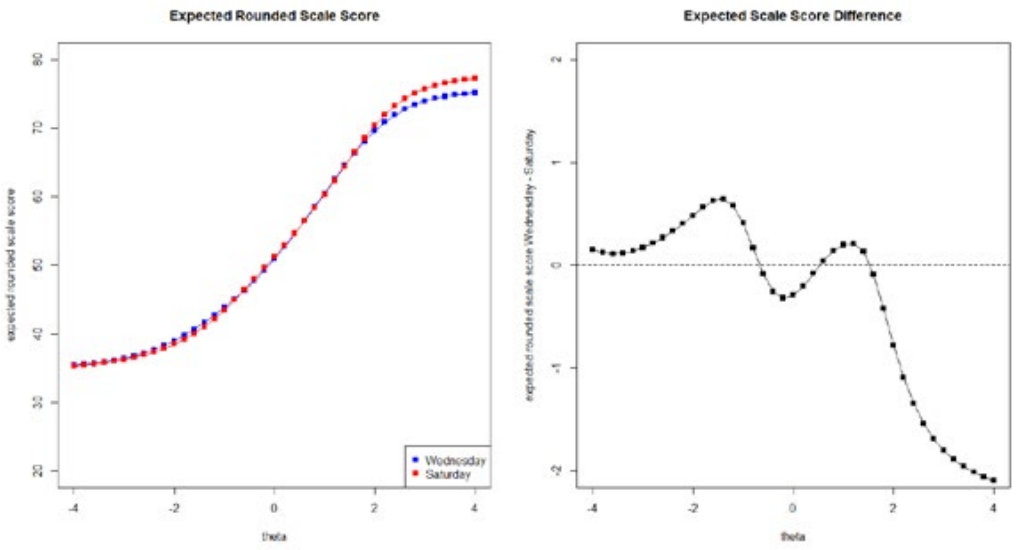


Figure 28.

First order equity using chained equipercentile equating for Writing.

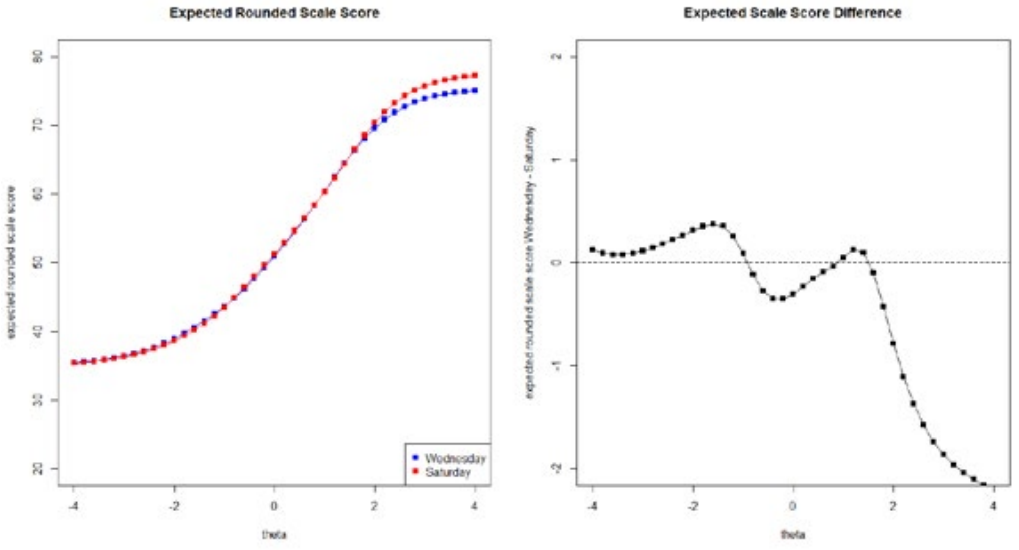


Figure 29.

Second order equity using IRT true score equating for Writing.

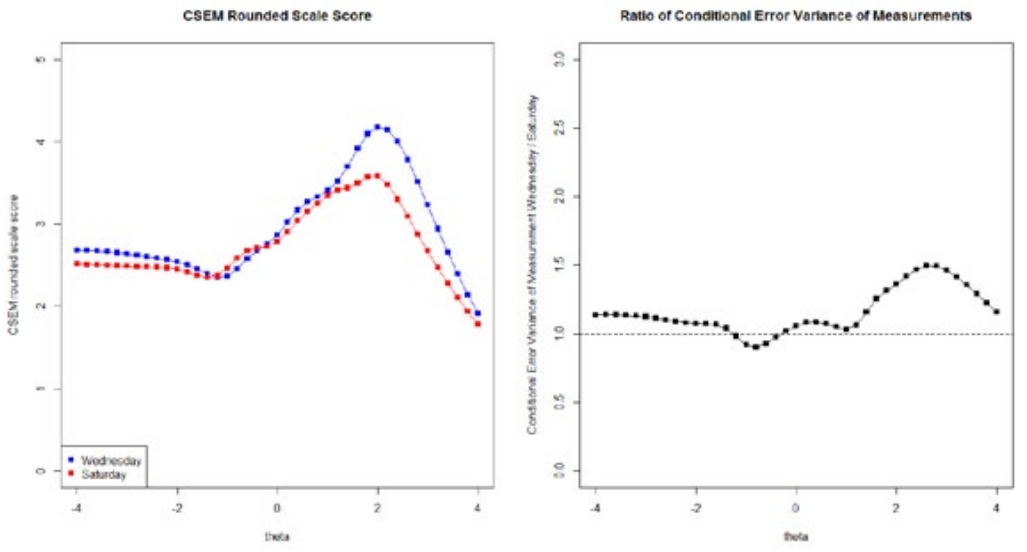


Figure 30.

Second order equity using IRT observed score equating for Writing.

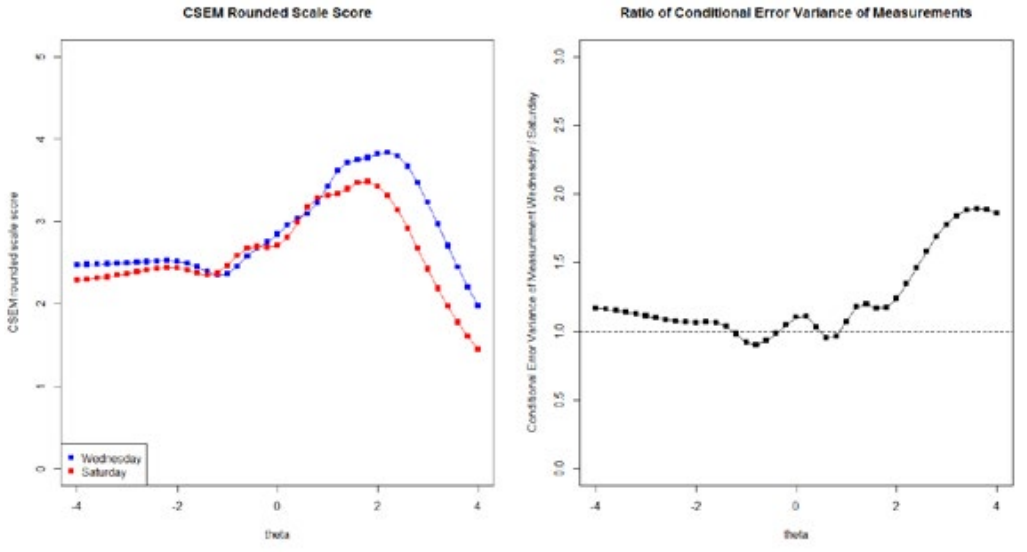


Figure 31.

Second order equity using frequency estimation equating for Writing.

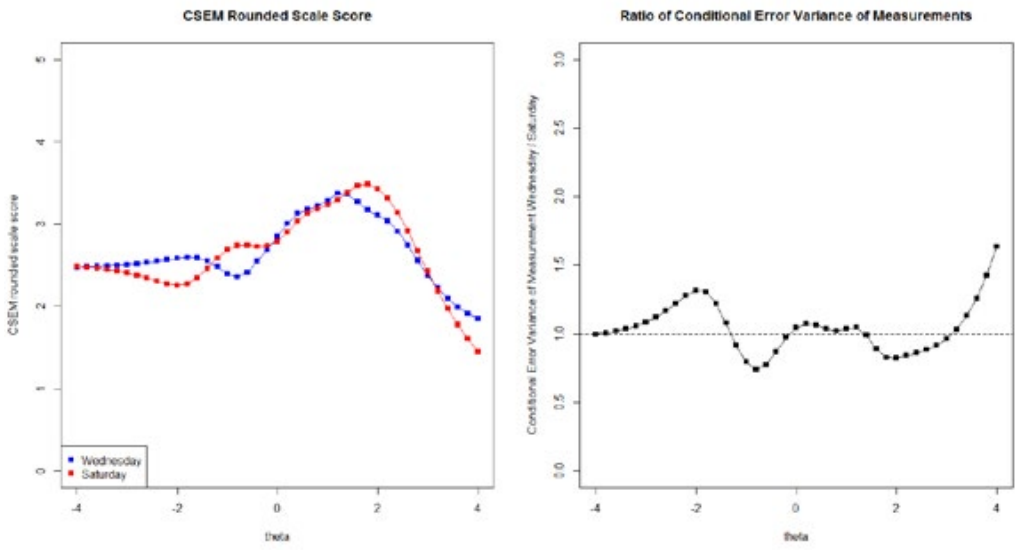


Figure 32.

Second order equity using chained equipercentile equating for Writing.

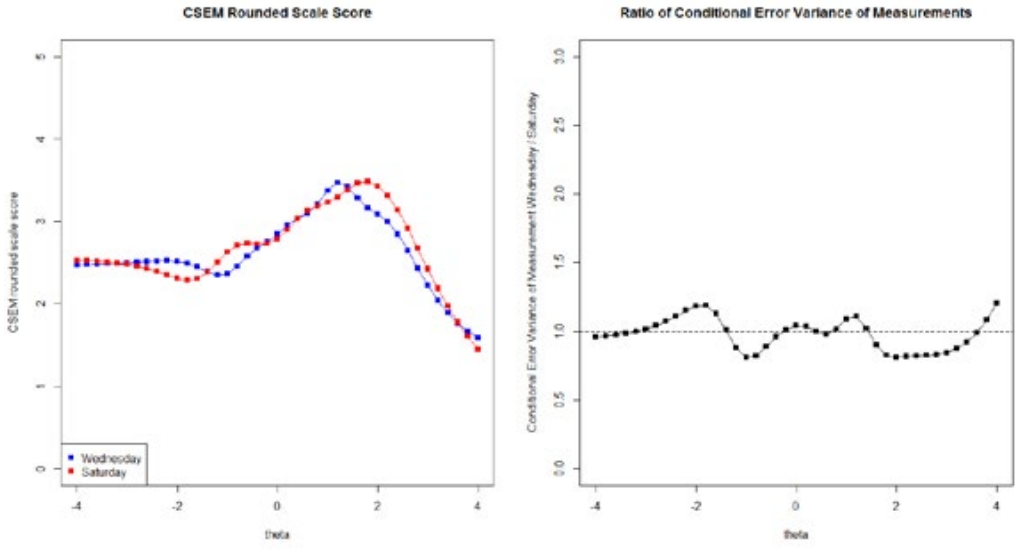


Figure 33.

Bootstrap test for equity for IRT true score equating for Writing

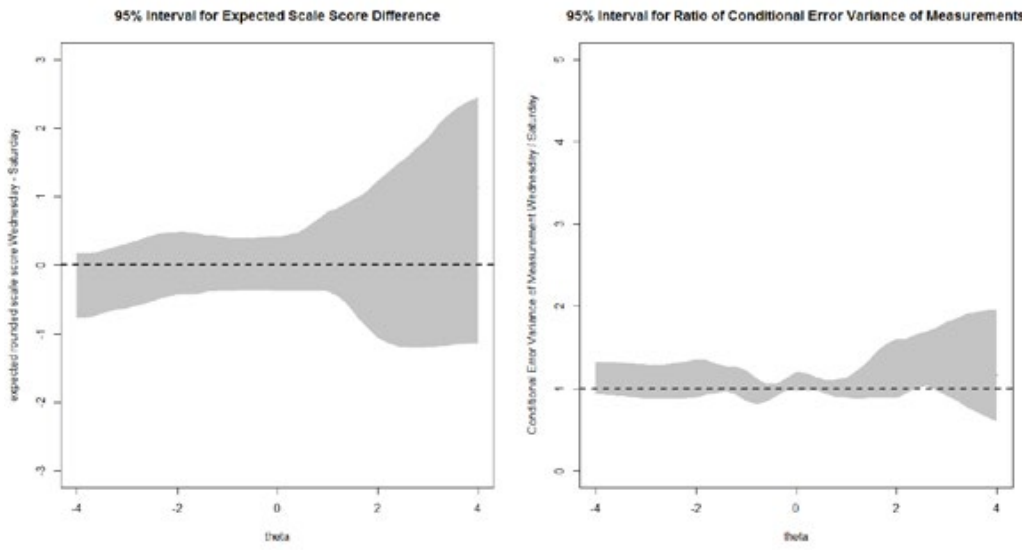


Figure 34.

Bootstrap test for equity for IRT observed score equating for Writing.

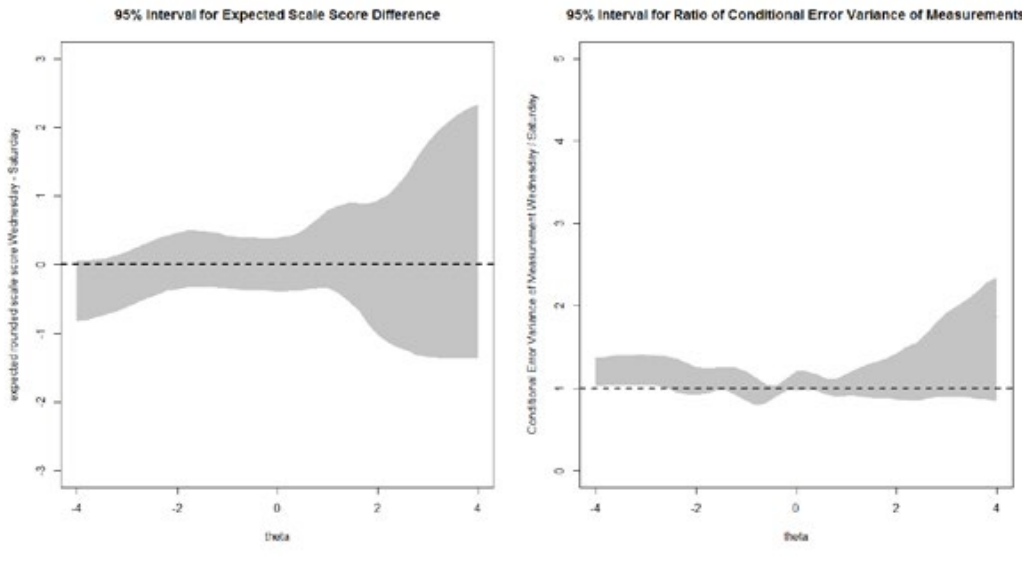


Figure 35.

Bootstrap test for equity for frequency estimation equating for Writing.

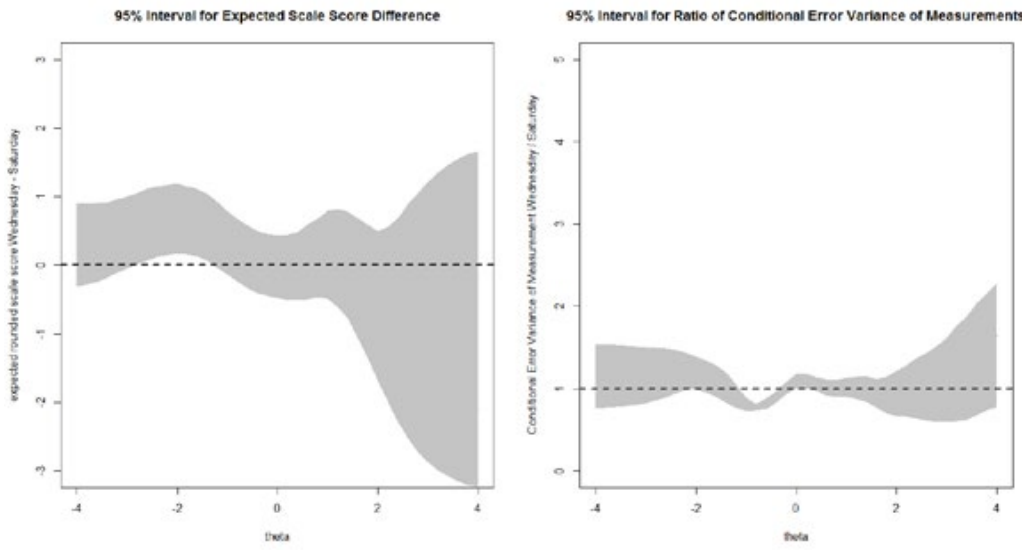
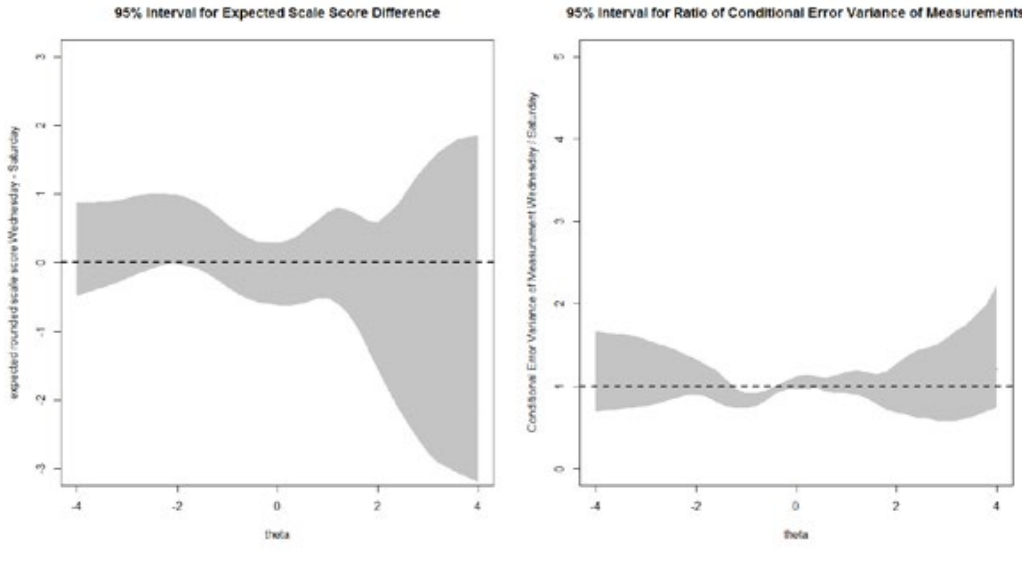


Figure 36.

Bootstrap test for equity for chained equipercentile equating for Writing.



In Table 3, at a given examinee ability level, each equating method has a value of zero if the 95% bootstrap confidence interval for testing the first-order equity includes zero and has a value of one if the confidence interval does not include zero. In Table 4, each equating method has a value of zero if the 95% bootstrap confidence interval for testing the second-order equity includes one while it has a value of one if the confidence interval does not include one. The last two rows provide a summary of the flags used to indicate whether the first-order equity or the second-order equity property was preserved at each examinee level. The first row contains the total number of flags, and the second row contains the number of flags where examinee ability level is higher than two ('Flag at ≥ 2.0'). It was decided to examine the examinee ability level of two or higher because the examinee ability level of two is equivalent to approximately 70 on the 20 to 80 pre-2015 PSAT/NMSQT scale, and this score range is usually important for the scholarship competition. Flags at greater than or equal to two appeared to vary depending on the equating method. These flags may help to decide which equating method will be selected since the score range is important for the PSAT/NMSQT. It is sensible to select an equating method that produces smaller numbers of flags at the score level that the assessment is most concerned about.

In practice, the discrepancy indices are often examined to assess multiple equating methods. A method providing smaller discrepancy indices is considered to be a better method in terms of preserving equity properties. However, smaller discrepancy indices may not provide sufficient information for a specific score level of interest. For example, for Critical Reading, the IRT true score equating method has the smallest D_1 index (Table 2). In Figure 1, the difference in the expected scale scores between the two forms seemed small. Thus, one might select the IRT true score equating as the final equating method. In fact, this result is consistent with the previous research findings that the IRT true score equating method preserves first-order equity compared to other equating methods (Kim et al., 2005; Tong & Kolen, 2005; Lee et al., 2010). However, the bootstrap test for first-order equity revealed that the confidence interval band for the examinee ability score between 2.0 and 3.0, which is important for the scholarship competition, did not include zero, indicating that the discrepancy between the two forms at that ability level is significant for IRT true score equating. With respect to the first-order equity at the local level, equating methods other than IRT true score equating should be used.

Limitations of the discrepancy index were also found for second-order equity. The D_2 index in Table 2 shows that for Math, overall, the FE and CE methods satisfied second-order equity better than the IRT methods. However, the confidence interval band for the examinee ability score at around 3.0 or higher did not include one for the FE and CE methods (Figures 23 and 24), while the confidence interval band for the same examinee ability level included one for the IRT true and observed score equating methods. Thus, the IRT methods preserved second-order equity for the higher examinee ability scores better than the FE or CE methods, even though overall the FE or CE methods preserved second-order equity better than the IRT methods.

Table 3.

Bootstrap Test for First Order Equity on Rounded Score

theta	CR				Math				Writing			
	IRT true	IRT obs	FE	CE	IRT true	IRT obs	FE	CE	IRT true	IRT obs	FE	CE
-4.0	0	0	0	0	0	0	0	0	0	0	0	0
-3.8	0	0	0	0	0	0	0	0	0	0	0	0
-3.6	0	0	0	0	1	0	0	0	0	0	0	0
-3.4	0	0	0	0	1	0	0	0	0	0	0	0
-3.2	0	0	0	0	1	0	0	0	0	0	0	0
-3.0	0	0	0	0	1	0	0	0	0	0	0	0
-2.8	0	0	0	0	1	0	1	0	0	0	1	0
-2.6	0	0	0	0	1	0	1	0	0	0	1	0
-2.4	0	0	1	0	1	0	1	0	0	0	1	0
-2.2	0	0	1	0	1	0	1	0	0	0	1	0
-2.0	0	0	1	0	0	0	1	1	0	0	1	0
-1.8	0	0	1	0	0	0	1	1	0	0	1	0
-1.6	0	0	1	1	0	0	1	1	0	0	1	0
-1.4	0	0	1	1	0	0	1	1	0	0	1	0
-1.2	0	0	1	0	0	0	1	0	0	0	0	0
-1.0	0	0	1	0	0	0	1	0	0	0	0	0
-0.8	0	0	0	0	0	0	1	0	0	0	0	0
-0.6	0	0	0	0	0	0	1	0	0	0	0	0
-0.4	0	0	0	0	0	0	0	0	0	0	0	0
-0.2	0	0	0	0	0	0	0	0	0	0	0	0
0.0	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	1	0	0	0	0	0
0.4	0	0	0	0	0	0	1	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0	0	0
1.0	0	0	0	1	0	0	0	0	0	0	0	0
1.2	0	0	1	1	0	0	0	0	0	0	0	0
1.4	0	0	1	1	0	0	0	0	0	0	0	0
1.6	0	0	1	1	0	0	0	0	0	0	0	0
1.8	1	1	1	1	0	0	0	0	0	0	0	0
2.0	1	1	1	1	0	0	0	0	0	0	0	0
2.2	1	1	0	0	0	0	0	0	0	0	0	0
2.4	1	0	0	0	0	0	0	0	0	0	0	0
2.6	1	0	0	0	0	0	0	0	0	0	0	0
2.8	1	0	0	0	0	0	0	0	0	0	0	0
3.0	0	0	0	0	0	0	0	0	0	0	0	0
3.2	0	0	0	0	0	0	0	0	0	0	0	0
3.4	0	0	0	0	0	0	0	0	0	0	0	0
3.6	0	0	0	0	0	0	0	0	0	0	0	0
3.8	0	0	0	0	0	0	0	0	0	0	0	0
4.0	0	0	0	0	1	0	0	0	0	0	0	0
Total Flag	6	3	13	8	9	0	14	4	0	0	8	0
Flag at >= 2.0	5	2	1	1	1	0	0	0	0	0	0	0

Table 4.

Bootstrap Test for Second Order Equity on Rounded Score

theta	CR				Math				Writing			
	IRT true	IRT obs	FE	CE	IRT true	IRT obs	FE	CE	IRT true	IRT obs	FE	CE
-4.0	0	0	0	0	0	0	0	0	0	1	0	0
-3.8	0	0	0	0	0	0	0	0	0	1	0	0
-3.6	0	0	0	0	0	0	0	0	0	1	0	0
-3.4	0	0	0	0	0	0	0	0	0	1	0	0
-3.2	0	0	0	0	0	0	0	0	0	1	0	0
-3.0	0	0	0	0	1	0	0	0	0	1	0	0
-2.8	0	0	0	0	1	0	0	0	0	1	0	0
-2.6	0	0	0	0	1	1	0	0	0	1	0	0
-2.4	0	0	0	0	1	1	0	0	0	0	0	0
-2.2	0	0	0	0	1	1	0	0	0	0	0	0
-2.0	0	1	0	0	1	1	1	1	0	0	0	0
-1.8	0	1	0	0	1	1	1	1	0	0	0	0
-1.6	1	0	0	0	1	1	1	1	0	0	0	0
-1.4	1	0	0	0	1	1	1	1	0	0	0	0
-1.2	1	0	1	0	1	1	1	1	0	0	0	1
-1.0	0	0	1	1	1	1	1	1	0	0	1	1
-0.8	0	0	1	1	1	1	1	1	0	0	1	1
-0.6	0	0	1	1	1	1	1	1	0	0	1	1
-0.4	0	0	0	0	1	1	1	1	0	0	1	0
-0.2	0	1	0	1	1	1	1	1	0	0	0	0
0.0	1	1	1	1	1	1	1	1	0	0	0	0
0.2	1	1	1	1	1	1	1	1	0	0	1	0
0.4	1	1	1	1	1	1	1	1	0	0	0	0
0.6	1	1	1	1	1	1	0	1	0	0	0	0
0.8	1	1	1	1	0	1	0	0	0	0	0	0
1.0	0	0	0	0	0	0	0	0	0	0	0	0
1.2	0	0	0	0	0	0	0	0	0	0	0	0
1.4	0	0	0	0	0	0	0	0	0	0	0	0
1.6	0	0	0	0	0	0	0	0	0	0	0	0
1.8	1	0	0	0	0	0	0	0	0	0	0	0
2.0	1	0	0	0	0	0	0	0	0	0	0	0
2.2	0	0	0	0	0	0	0	0	0	0	0	0
2.4	0	0	0	0	0	0	0	0	1	0	0	0
2.6	0	0	0	0	0	0	0	0	1	0	0	0
2.8	0	0	0	0	0	0	0	0	0	0	0	0
3.0	0	0	0	0	0	0	0	0	0	0	0	0
3.2	0	0	0	0	0	0	0	1	0	0	0	0
3.4	0	0	0	0	0	0	1	1	0	0	0	0
3.6	0	0	0	0	0	0	1	1	0	0	0	0
3.8	0	0	0	0	0	0	1	1	0	0	0	0
4.0	0	0	0	0	0	0	1	1	0	0	0	0
Total Flag	10	8	9	9	19	18	17	19	2	8	5	4
Flag at >= 2.0	1	0	0	0	0	0	4	5	2	0	0	0

Conclusion

For test fairness, equating practices should ensure that test scores from different administrations are equivalent. The current study provides a useful tool for the local evaluation of the equity of test scores. The current study shows the limitations of the discrepancy indices that are often used to assess the adequacy of an equating method, and suggests that bootstrap intervals can be usefully used as an alternative method to assess the adequacy of the various equating methods. The approach is particularly helpful for high-stakes assessments where the determination of cut scores takes on great importance and has implications for practice.

Although the results of the bootstrap test revealed great utility to assess equity properties, it is also important to notice that the results are limited to the specific sample data from the pre-2015 PSAT/NMSQT test administration in which the ability difference between new and old forms was relative large due to the difference in the target populations. In addition, the equating design for the current study was the common nonequivalent design. Thus, simulation studies that include different equating designs and different levels of examinee ability differences (between old and new forms) should be examined in future studies. Sensitivity analysis can be conducted to assess the efficacy of the bootstrap test.

References

Boos, D. D., & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics*, *31*(1), 69–82.

Cai, L. (2013). flexMIRT (Version 2): Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cook, L. L., Dunbar, S. B., & Eignor, D. R. (1981, April). IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at the meeting of the American Educational Research Association, Los Angeles, CA

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.

Kim, D. I., Brennan, R. L., & Kolen, M. J. (2005). A comparison of IRT equating and beta 4 equating. *Journal of Educational Measurement*, *42*(2), 77–99.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*(4), 285–307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129–140.

Lee, E., Lee, W.-C., & Brennan, R. L. (2010). *Assessing equating results based on first-order and second-order equity*. (CASMA Research Report No. 31). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *8*, 453–461.

Morris, G. M. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). New York: Academic Press.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement*, *29*(6), 418–432.

The Research department actively supports the College Board's mission by:

- Providing data-based solutions to important educational problems and questions
- Applying scientific procedures and research to inform our work
- Designing and evaluating improvements to current assessments and developing new assessments as well as educational tools to ensure the highest technical standards
- Analyzing and resolving critical issues for all programs, including AP®, SAT®, PSAT/NMSQT®
- Publishing findings and presenting our work at key scientific and education conferences
- Generating new knowledge and forward-thinking ideas with a highly trained and credentialed staff

Our work focuses on the following areas

Admission	Measurement
Alignment	Research
Evaluation	Trends
Fairness	Validity

Follow us online: research.collegeboard.org

CollegeBoard

CollegeBoard