RESEARCH REPORT 2012-4

CollegeBoard

Exploring Equity Properties in Equating Using AP® Examinations

By Eunjung Lee, Won-Chan Lee, and Robert L. Brennan

Eunjung Lee is a research assistant in educational measurement and statistics at the College of Education, University of Iowa.

Won-Chan Lee is associate professor and associate director for the Center for Advanced Studies in Measurement and Assessment (CASMA).

Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and director for the Center for Advanced Studies in Measurement and Assessment (CASMA).

Mission Statement

The College Board's mission is to connect students to college success and opportunity. We are a not-for-profit membership organization committed to excellence and equity in education.

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.

RESEARCH

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

Contents

Figures

Executive Summary

In almost all high-stakes testing programs, test equating is necessary to ensure that test scores across multiple test administrations are equivalent and can be used interchangeably. Test equating becomes even more challenging in mixed-format tests, such as Advanced Placement Program® (AP®) Exams, that contain both multiple-choice and constructed response items. This report examines (1) the performance of various equating methods in terms of first- and second-order equity properties using mixed-format tests; (2) the effect of underlying psychometric models on the assessment of the performance of the equating methods; and (3) the relationship between reliability and equity properties in equating. Three AP Exams (Biology, English Language and Composition, and French Language and Culture) were analyzed with the common-item, nonequivalent-groups design. The 11 equating methods were analyzed, and the results were obtained and compared based upon two different psychometric model frameworks: the two-parameter beta binomial and item-response theory (IRT). In general, the results showed that the performance of various equating methods in terms of equity properties depended on the psychometric model assumed. Furthermore, this report provides empirical evidence that the magnitude of reliability plays a role in achieving the equity properties for the various equating methods.

Introduction

When equating is performed properly, scores on alternate forms can be used almost interchangeably. Among many desirable properties of equating, Lord (1980) proposed the equity property (p. 195). Lord's equity property is applicable if the observed-score distribution on the old form is the same as the distribution of new-form equated scores for examinees with a given true score. Lord also showed, however, that it is impossible to satisfy the equity property unless the two forms are essentially identical, in which case equating is unnecessary. As a practical solution to this contradiction, Morris (1982) suggested a "weak" definition of equity, which applies if examinees with a given true score have the same expected score on the old form and the new equivalent form. Morris's weak definition of equity is often referred to as first-order equity (FOE). So-called second-order equity (SOE) requires the same variance of the observed score distribution for the two forms, depending upon a given true score. In the current measurement literature, these two equity properties together are often called weak equity and considered when evaluating or comparing various equating procedures.

In the measurement literature, there have been several studies that focus on the comparison and evaluation of different types of equating methods that are based on different equating criteria, including FOE and SOE (Han, Kolen, & Pohlmann, 1997; Kim, Brennan, & Kolen, 2005; Tong & Kolen, 2005; Wang, Hanson, & Harris, 2000). It should be noted, however, that there have been research gaps in the equity property area. First, a very limited number of types of equating designs and test formats have been studied. In terms of equating designs, all the previous studies have been conducted based only on the random-groups equating design. In terms of test formats, it is hard to find a study that used a mixed-format test for comparing various equating results in terms of FOE and SOE. Thus, different equating designs and test formats need to be examined in order to broaden and build our knowledge of the equity property.

Second, there have been only a few studies that examine the effect of underlying psychometric models on the assessment of the performance of equating methods in terms of FOE and SOE (Kim et al., 2005). To examine whether an equating method attains FOE and SOE, expected scores and conditional standard errors of measurement (CSEMs) are computed under certain assumptions of a psychometric model. Accordingly, it can be said that the extent to which a given equating method satisfies FOE and SOE is likely to be determined by the underlying psychometric framework. Most of the previous studies used an IRT framework (Lee, Lee, & Brennan, 2010; Tong & Kolen, 2005). However, other types of psychometric frameworks have been rarely used; subsequently, little is known about their effect on the equity property of various equating methods.

Third, the relationship between reliability and equity properties in equating has not yet been examined. High reliability is a ubiquitous requirement for educational and psychological tests. Ironically, high-reliability is rarely considered as a requirement for equating, partly because most equating procedures are not directly associated with a true-score test theory. Consequently, the role of reliability in equating has not been firmly established, and its effect on equating results is relatively unknown. Recently, Brennan (2010) suggested a theoretical foundation for the relationship between reliability and FOE and SOE. However, a research study that uses a real data set to support Brennan's study has not yet been conducted. Therefore, this study attempts to fill these gaps by achieving three research objectives, as noted below.

Objectives

The objectives of this study are threefold: (1) To assess the performance of various equating methods in terms of FOE and SOE using mixed-format tests with the common-item, nonequivalent-groups design; (2) to examine the effect of underlying psychometric models on the assessment of the performance of equating methods; and (3) to investigate the relationships between test reliability and equity properties in equating.

To achieve the goals of this study, three sets of data were analyzed. First, the intact forms of the three AP Exams were analyzed to examine FOE and SOE in mixed-format tests. Second, to investigate the effect of underlying psychometric models on the equating results, pseudo–test forms that consisted solely of the multiple-choice (MC) items on the AP Biology Exam were constructed and analyzed. Third, to examine the relationships between reliability and FOE and SOE, pseudo-test forms were constructed by shortening the MC items on the AP Biology Exam.

In order to assess various equating methods in terms of FOE and SOE, expected scores and CSEMs are computed under the assumption of a psychometric model. For the analysis of the intact forms of AP Exams, a unidimensional IRT framework was used. For the analysis of the tests with MC items only, the results based on the two-parameter beta binomial (BB) framework were compared with those based on the IRT framework.

This study assesses the performance of 11 equating methods: Tucker method (Gulliksen, 1950), Levine observed-score method (Levine, 1955), Levine true-score method (Levine, 1955), unsmoothed chained equipercentile method (Angoff, 1971), chained equipercentile method with log-linear presmoothing, chained equipercentile method with cubic-spline postsmoothing, unsmoothed frequency estimation method (Angoff, 1971), frequency estimation method with log-linear presmoothing, frequency estimation method with cubic-spline postsmoothing, IRT true-score method (Lord, 1980), and IRT observed-score method (Kolen, 1981; Lord, 1980).

Relevant Research

FOE and SOE

Previous researchers have described procedures that can be used to compute conditional expected scores and CSEMs based on different psychometric models. Kolen, Hanson, and Brennan (1992) present a procedure under a strong true-score model. Kolen, Zeng, and Hanson (1996) describe a procedure using a dichotomous IRT model. Wang et al.'s (2000) procedures can be implemented with polytomous IRT models. Brennan (2010) provides an extensive discussion about equity under classical test theory as well as computational formulas for FOE and SOE with the two-parameter BB model.

There are several studies that have used these procedures in examining FOE and SOE. For example, Tong and Kolen (2005) compared the performance of three equating methods (the equipercentile method with log-linear presmoothing, IRT true-score method, and IRT observed-score method) in terms of FOE and SOE using Kolen et al.'s (1996) dichotomous IRT procedure. They found that the IRT true-score method outperformed the other two methods in terms of preserving FOE, and that both the IRT observed-score equating method and the equipercentile method performed better than the IRT true-score method in preserving SOE.

Kim et al. (2005) compared four equating methods (the IRT true- and observed-score equating methods, and beta 4 true- and observed-score equating methods) in terms of FOE and SOE under two psychometric models (the 3PL IRT model and the beta 4 model). Their results showed that regardless of the psychometric model assumed, the true-score equating methods better preserved FOE than the observed-score equating methods, whereas the observed-score equating methods better satisfied SOE than the true-score equating methods. They further showed that the IRT true-score equating method satisfied FOE better than the beta 4 true-score equating method when the true-score distribution was estimated under the 3PL IRT model. The opposite result was found when the beta 4 model was assumed.

Recently, Lee et al. (2010) compared the results of seven equating methods based on FOE and SOE using Kolen et al.'s (1996) dichotomous IRT procedure. They used both normal and uniform weights in computing overall indices to examine whether an equating method satisfies FOE or SOE. Their results showed that the IRT true-score method preserved FOE better than any other method, regardless of weighting. When normal weights were used, the IRT observed-score method was found to be the most likely method to satisfy SOE, and when uniform weights were used, smoothed equipercentile equating using the cubic-spline postsmoothing was found to be the most likely method to preserve SOE.

All the equating designs used in these previous studies were based on the random-groups equating designs, and tests with only MC items were used. In this study, we compare various equating results for the mixed-format tests with the common-item, nonequivalent-groups design (Kolen & Brennan, 2004).

Role of Reliability in FOE and SOE

Recently, Brennan (2010) studied FOE and SOE for true-score and observed-score equating under certain assumptions of classical test theory. With this study, Brennan is perhaps the first who considered explicitly the role of reliability as it relates to the FOE and SOE in equating. Brennan derived results for both linear and curvilinear equating, and a summary is provided below:

For linear equating,

1. FOE is satisfied for applied true-score equating[1];
2. FOE is satisfied for observed-score equating, if the reliabilities for the two forms are equal;
3. SOE for applied true-score equating is satisfied, if
    a. for all true-score levels, the ratio of conditional error variances is a constant equal to the ratio of true-score variances, or
    b. the reliabilities for the two forms are equal *and* for each form, the conditional error variances are homogeneous; and
4. SOE for observed-score equating is satisfied, if
    a. for all true-score levels, the ratio of conditional error variances is a constant equal to the ratio of observed-score variances, or

1. The term "applied true-score equating," used in Brennan (2010), refers to the linear true-score equating where the new-form true score is replaced by the new-form observed score.

   b. the reliabilities for the two forms are equal *and* for each form, the conditional error variances are homogeneous.

For curvilinear equating,

1. FOE for applied true-score equating is more nearly satisfied, if the reliability for the new form is high;
2. FOE for observed-score equating is more nearly satisfied under the condition of equal and high reliabilities for the old and new forms;
3. SOE for applied true-score equating is more nearly satisfied, if
   a. for all true-score levels, the ratio of conditional error variances is a constant equal to the ratio of true-score variances *and* the reliabilities for both forms are high, or
   b. the reliabilities for the two forms are equal, the conditional error variances are homogeneous, *and* the reliabilities for both forms are high; and
4. SOE for observed-score equating is more nearly satisfied, if
   a. for all true-score levels, the ratio of conditional error variances is a constant equal to the ratio of observed-score variances *and* the reliabilities for both forms are high, or
   b. reliabilities for the two forms are equal, the conditional error variances are homogeneous, *and* the reliabilities for both forms are high.

The summary presented above suggests that the magnitude of reliability plays a role in achieving FOE and SOE for the curvilinear equating procedures.

Methods

Data

Equity Properties for Mixed-Format Tests

The data used in this study were from three AP Exams: Biology, English Language, and French. Altogether there are four linkages: two linkages for Biology, one for English Language, and one for French. The equating design was the common-item, nonequivalent-groups design (Kolen & Brennan, 2004). The original AP Exams are the mixed-format tests that contain both MC items and constructed-response (CR) items. Operationally, the MC items in the AP Exams are formula scored, and noninteger weights are assigned to each MC and CR section so that the weighted composite scores across the two forms are equal. However, in this study, number-correct scoring was employed for the MC items. In addition, integer weights were assigned to items, and thus the weighted total score for the MC items and the weighted total score for the CR items were often different across forms. Note that because the original AP data were modified, results and findings from this study should not be generalized to the operational AP Exams.

Table 1 presents the integer weights assigned to the MC and CR items, and the contributions of each item format to a composite score. Composite scores in Table 1 are weighted summations of the number-correct scores on the MC items and the summed scores on the CR items. Table 2 provides descriptive statistics for the weighted composite score as well as

the number of examinees actually included in this study. These sample sizes were calculated after eliminating examinees who answered fewer than 80% of the MC questions. The Pearson correlations between the composite scores and the common-item (CI) scores were quite high for all the forms (all of them were above .75).

In converting the composite raw scores to scale scores, scale scores developed for the AP redesign and equating research were used. The scale scores were found by rounding and then normalizing the weighted composite scores. The scale scores were rounded to integers and truncated so that they were between 0 and 70. The mean of the scale scores was approximately 35, and the standard deviation was approximately 10.

Effects of Psychometric Models on Equity Properties

To examine whether different psychometric models would yield different results for expected scores and CSEMs, we constructed pseudo-test forms using four AP Biology Exams from 2004, 2005, 2006, and 2007. These pseudo-forms were constructed simply by eliminating all the CR items in each test so that each test consisted of the MC items only. Thus, these pseudo-forms are called full-length MC pseudo-tests, for descriptive purpose, in this paper. In analyzing these pseudo-forms, unweighted summations of the number-correct scores of the MC items were used as the total raw scores. The scale scores were developed by conducting an unsmoothed equipercentile equating from the intact form (a full-length test with the MC and CR items) to the pseudo-form (the same test with the MC items only) with a single-groups design. Table 3 provides descriptive statistics for unweighted number-correct scores of the MC and CI items for the full-length MC pseudo-tests. Reliabilities were quite high for all the forms (alpha coefficients for all tests were above .93).

Relationship Between Reliability and Equity Properties

Because the reliabilities of all the AP Exams were high (when we considered the MC items only, the alpha coefficients for all the forms used in this study were above .9), and there were no significant differences in the reliabilities among the tests, we constructed other pseudo-forms to examine the relationships between test reliability and FOE and SOE. These pseudo-forms were constructed using two AP Biology Exams, from 2004 and 2006. For this analysis, only the MC items were used.

Pseudo-tests for three reliability conditions were constructed as follows. For each form, the items in the full-length MC pseudo-test, excluding the common items, were rank ordered by their discrimination levels (i.e., point-biserial correlations), and the first 36 highly correlated items were eliminated. Then the same procedures were applied to the common items, and the first 13 highly correlated common items were eliminated. Finally, the rest of the items from these two sets of items were merged to create the low-reliability test. For the medium-reliability test, the 19th to 54th items based on the rank-ordered distribution were eliminated from the noncommon items, and the seventh to 19th items were eliminated from the common items. Then the rest of the items were merged to construct the medium-reliability test. Finally, to create the high-reliability test, the last 36 items were eliminated from the noncommon items, and the last 13 items were eliminated from the common items. For all three reliability conditions, there was a total of 50 items for Biology 2004 and 49 items for Biology 2006. These tests are called shortened MC pseudo-tests in this paper.

We examined the relationships between reliability and FOE and SOE by comparing the results from each reliability condition. In analyzing these pseudo-forms, unweighted summations of the number-correct scores of the MC items were used as the total raw scores. Scale scores

were developed by conducting unsmoothed equipercentile equating from the intact form to the pseudo-form with a single-groups design.

Descriptive statistics for unweighted number-correct scores of the MC and CI items for the shortened MC pseudo-tests are presented in Table 4. The difference in the alpha coefficient between the high-reliability condition and the medium-reliability condition for the 2004 form was .043; for the 2006 form, the difference was .039. The difference in the alpha coefficient between the medium-reliability condition and the low-reliability condition for the 2004 form was .049; for the 2006 form, the difference was .044. Thus, the difference in the alpha coefficient between the high-reliability condition and the low-reliability condition for the 2004 form was .092; for the 2006 form, the difference was .084.

Equating Methods

To achieve the goals of this study, 11 equating methods were considered:

a. Tucker method

b. Levine observed-score method

c. Levine true-score method

d. Unsmoothed chained equipercentile method

e. Chained equipercentile method with log-linear presmoothing

f. Chained equipercentile method with cubic-spline postsmoothing

g. Unsmoothed frequency estimation method (sometimes called the post-stratification method)

h. Frequency estimation method with log-linear presmoothing

i. Frequency estimation method with cubic-spline postsmoothing

j. IRT true-score method

k. IRT observed-score method

The first three methods are linear equating methods, while the other eight are curvilinear methods. For the log-linear presmoothing methods, the fixed values of the presmoothing parameter pairs were (6, 6, 1 : 6, 6, 1), which means that for both old and new forms, six moments in the marginal distribution of the composite score and common-item score as well as the first cross-moment in the bivariate distribution were preserved. Smoothing parameters for the cubic-spline postsmoothing method were chosen using a judgmental procedure based primarily on whether or not smoothed scores were within plus or minus one raw score standard error of equating between percentile ranks of 0.5 and 99.5. For all the tests, the postsmoothing parameter $S = 0.1$ met the criteria for parameter selection reasonably well. Figures 1 and 2 provide examples of the postsmoothed and unsmoothed equating relationship for AP Biology 2004–2006 with plus and minus one standard error of equating. The conditional standard errors of equating were calculated using a bootstrap procedure with 1,000 replications. The smoothed equivalents for $S = 0.1$ are much smoother than the unsmoothed equivalents, but the smoothed results follow the unsmoothed equivalents quite closely.

For the IRT equating methods, IRT calibration was done for the old form and the new form separately, assuming the three-parameter logistic (3PL) model (Birnbaum, 1968) for the MC items and the graded response (GR) model (Samejima, 1969) for the CR items. After item parameters and posterior proficiency distributions of both forms were obtained, the Stocking–Lord method (Stocking & Lord, 1983) was used to transform the estimated item parameters and the quadrature points of the posterior distribution on the new-form scale to the old-form scale.

Estimation Methodology

When equating methods are applied to real data, an estimate of the true-score distribution is required to assess FOE and SOE. This is because these two properties refer to the relationship between the two conditional distributions of observed scores given their true scores (Kim et al., 2005). Previous researchers have described several procedures that can be used to compute conditional expected scores and CSEMs based on different psychometric models. For example, Kolen, Hanson, and Brennan (1992) provided a general approach for estimating CSEMs for scale scores and considered an application of a strong true-score model. Kolen, Zeng, and Hanson (1996) presented specifics on how to use Kolen et al.'s (1992) general approach in a dichotomous IRT framework. In this study, both Kolen et al.'s (1992) framework of strong true-score model and Kolen et al.'s (1996) IRT model approaches were used.

Under the BB framework, the probability that a raw score random variable X is equal to i (i = 0,1 … , k) on a k-item test is,

$$\Pr(X = i) = \int_0^1 \Pr(X = i \mid \tau) g(\tau) d\tau, \quad (1)$$

where τ is the examinee's true proportion of items correct. The true-score distribution, $g(\tau)$, is assumed to belong to the two-parameter beta family of distributions. The conditional error distribution, $\Pr(X = i \mid \tau)$, is assumed to be binomial and is expressed as follows:

$$\Pr(X = i \mid \tau) = \binom{k}{i} \tau^i (1-\tau)^{k-i}. \quad (2)$$

A similar procedure can be used for a unidimensional IRT framework. Under IRT, ability θ serves as the conditioning variable instead of τ under the strong true-score model. The conditional distribution of number-correct raw scores, symbolized as $\Pr(X = i \mid \theta)$, can be modeled using a compound binomial model. The Lord and Wingersky (1984) recursion formula is typically used for computing the conditional raw-score distribution based on item parameter estimates.

Then a raw-to-scale score transformation can be applied to the conditional distribution of number-correct raw scores at a given ability (or a true score) in order to produce the conditional probability distribution of scale scores. The mean of this conditional distribution is the true (expected) scale score at that ability level, which is given by

$$\xi(\tau) = Exp[s(X) \mid \tau] = \sum_{i=0}^{k} s(X) \Pr(X = i \mid \tau) \quad (4)$$

where Exp refers to the expected value, and the raw-to-scale score transformation is symbolized as s. The standard deviation of this conditional scale score distribution at a given ability level is taken to be the CSEM at that ability, which is

$$\sigma[s(X) \mid \tau] = \sqrt{EV \mid \tau} = \sqrt{Exp\left\{[s(X) - \xi(\tau)]^2 \mid \tau\right\}}, \quad (5)$$

where $EV|\tau$ refers to the conditional error variance of measurement at a given ability (or a true score). Under the IRT framework, θ is replacing τ everywhere it appears in Equations (4) and (5).

In this study, we adopted a unidimensional IRT framework for assessing equity for the intact forms of the AP Exams, which are mixed-format tests consisting of MC and CR items. For the full-length and shortened MC pseudo-tests, we compared the results based on the BB and those based on the IRT framework. When we used the IRT framework, the true-score (θ) distribution was assumed to be a normal distribution with a mean of 0 and a standard deviation of 1. Furthermore, we assumed that the true-score (τ) distribution under the BB framework was a two-parameter beta distribution with both parameters set to 2 such that the distribution was symmetric.

Under the IRT framework, it was assumed that the true-score relationship between the two forms was curvilinear, and the relationship was defined according to the IRT true-score equating. On the other hand, under the BB framework, it was assumed that the true scores on the two forms were linearly related, and that the linear relationship was defined by Levine true-score equating. Note that both the IRT true-score equating and the Levine true-score equating are applied true-score equating methods whereby the new-form true scores are replaced by the new-form observed scores.

Evaluation Criteria

To demonstrate empirically each equating method's adequacy of preserving FOE and SOE, the differences in expected scores and CSEMs were plotted, and the overall discrepancy indices were computed. The present study adopted the same discrepancy indices as the ones used in Lee, Lee, and Brennan (2010):

$$D_1 = \frac{\sum_i w_i \left| \xi_{Y(\tau_i)} - \xi_{X(\tau_i)} \right|}{\sum_i w_i}, \tag{6}$$

and

$$D_2 = \sqrt{\frac{\sum_i w_i \left| (EV_Y | \tau_i) - (EV_X | \tau_i) \right|}{\sum_i w_i}}. \tag{7}$$

In Equation (6), $\xi_{Y(\tau_i)}$ refers to the expected scores of the old form at a given true score (ability level), while $\xi_{X(\tau_i)}$ refers to those of the new form[2], and w_i refers to the weight of τ_i. In Equation (7), $EV_Y|\tau_i$ refers to the conditional error variance of measurement of the old form at a given true score (ability level), and $EV_X|\tau_i$ refers to those of the new form. For the weight, both weights from a normal distribution and weights from a uniform distribution were used in computing D_1 and D_2. A discrete quadrature distribution with 31 quadrature points was used to compute D_1 and D_2; θ ranged from -3 to +3, and τ ranged from .125 to .875.

Computer Programs

IRT calibration was conducted using MULTILOG (Thissen, Chen, & Bock, 2003). For IRT equating, the 3PL model was used to estimate item parameters for the MC items, and the graded-response model (GRM) was used for the FR items. All equating procedures were performed using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). Expected scale scores and conditional standard errors of measurement under the IRT model were computed using POLYCSEM (Kolen, 2004).

2. *X*s in Equations (6) and (7) should be *eqY*(*x*) to more accurately represent the scores on the new form converted to the scale of the old form using an equating function. However, for notational simplicity, we used *X* instead of *eqY*(*x*).

Results[3]

Equity Properties for Mixed-Format Tests

As mentioned earlier, a unidimensional IRT framework was used to analyze the intact forms. Figures 3 to 6 show the differences in conditional expected composite raw scores. The differences in conditional expected scale scores are presented in Figures 7 to 10. In each plot, the horizontal axis represents expected scores on the old form, and the vertical axis represents the differences between the expected scores on the old form and those yielded by equivalents from each equating method. If FOE had held perfectly, the curves in these graphs would have been coincident with the horizontal zero line, meaning that the expected scores yielded by the equivalents were the same as the expected scores for the old form. The curves for the IRT true-score method preserved FOE throughout the whole score range for both the raw and scale scores, which is consistent with previous research findings (Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005). IRT observed-score methods for all the tests also seemed to preserve FOE quite well for both the raw and scale scores. For AP Biology Exams (see Figures 3, 4, 7, and 8), the curves for the linear equating methods were close to each other, those for the IRT equating methods did not depart from each other, and those for the traditional[4] curvilinear equating methods were grouped together. It seems that the linear equating methods preserved FOE better than the traditional curvilinear equating methods, especially for AP Biology 2005–2007. For the other two exams, especially for AP French, the curves for all the non-IRT-based equating methods overlapped considerably (see Figures 5, 6, 9, and 10), indicating that these methods yielded no meaningful differences in the extent to which FOE held.

The overall FOE indices (D_1) for all the tests are presented in Table 5. The overall FOE index (D_1) reports how well each equating method satisfies FOE. The lower the D_1 value, the better FOE was satisfied. Entries in italics and bold in Table 5 indicate an equating method that yields the lowest D_1 value for each equating linkage. The IRT true-score equating method preserved FOE better than any other method, regardless of weighting used to compute the D_1 statistic. This is consistent with the previous research findings (Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005) that the IRT true-score method preserves FOE well under the random-groups design. The differences in the D_1 values between the IRT true- and observed-score equating methods were quite small, which indicates that the IRT observed-score equating method also satisfies FOE. The D_1 values for the non-IRT equating methods for all the tests did not seem to differ much. These findings are consistent with the graphical representation of the results (Figures 3 to 10).

With regard to SOE, the differences in CSEMs for raw scores are illustrated in Figures 11 to 14. The differences in CSEMs for scale scores are depicted in Figures 15 to 18. In each plot, the horizontal axis represents expected scores on the old form, and the vertical axis represents the differences between the CSEM on the old form and those yielded by equivalents from each equating method. If SOE had held perfectly, there would have been no difference between the CSEMs on the two forms at each proficiency level, and the curves in each plot would have been a horizontal zero line. The pattern of findings for SOE, however, was less consistent than that for FOE. For AP English Language and Composition, the Tucker and Levine true-score methods, relative to the other methods, seemed to yield curves closest

3. Note again that results and findings from this study should not be generalized to the operational AP Exams because the original AP data were modified, and different scoring, weighting scheme, and scaling were used for this research.

4. In this paper, we categorized all four non-IRT-based curvilinear equating methods as traditional curvilinear equating methods for the descriptive purpose.

to the horizontal zero line (see Figures 13 and 17). For AP French, the SOE curves from all the equating methods did not depart much from each other (see Figures 14 and 18). For AP Biology 2004–2006 and 2005–2007 (see Figures 11, 12, 15, and 16), the curves for the IRT equating methods were gathered together, those for the linear equating methods were close to each other, and those for the traditional curvilinear equating methods went together. However, for the two AP Biology equating linkages, it is difficult to tell which equating method outperformed other equating methods in terms of preserving SOE from observing these graphs.

The overall SOE index (D_2) values for all the tests are presented in Table 6. The overall SOE index (D_2) empirically quantifies the overall differences in CSEMs between the old and new forms across all the proficiency levels. The lower the D_2 value, the better SOE was satisfied. Between the two IRT equating methods, IRT observed-score equating tended to have slightly lower D_2 values than the IRT true-score equating. This is consistent with the previous research findings (Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005) that IRT observed-score equating preserves SOE better than IRT true-score equating under the random-groups design. Otherwise, it does not appear that the D_2 values of one method were consistently lower than those of the other method. Therefore, it is difficult to conclude which method performs better in terms of preserving SOE.

Effects of Psychometric Models on Equity Properties

In analyzing the full-length MC pseudo-tests to examine whether different psychometric models would yield different results for FOE and SOE, the BB and IRT frameworks were fitted to the same equating data.

Figures 19 to 22 show the differences in conditional expected composite raw score. The differences in conditional expected scale score are presented in Figures 23 to 26. When the BB framework is used (see Figures 19, 21, 23, and 25), the horizontal axis in each plot represents expected scores on the old form, and the vertical axis represents the differences between the expected scores on the old form and those yielded by equivalents from each equating method. It is assumed that two true scores for the old and new forms are linearly related, and the relationship is determined by the Levine true-score equating. The Levine true-score equating method satisfied FOE, which is shown in Figures 19, 21, 23, and 25. This result is consistent with Brennan's (2010) and Hanson's (1991) findings, which showed that FOE is satisfied for the applied true-score equating when the equating relationship is linear. The Levine observed-score equating method also satisfied FOE throughout the whole score range when the BB model was assumed (Figures 19, 21, 23, and 25). Considering the quite similar reliabilities for the two forms (the differences between alpha coefficients for the two forms were .001 for both of the linkages), we expected this finding based upon Brennan's (2010) findings, which showed that when the equating relationship is linear, FOE is satisfied for observed-score equating if the reliabilities for the two forms are equal. For AP Biology 2004–2006, the Tucker equating method also seemed to preserve FOE throughout the whole score ranges, and the curves for all the curvilinear equating methods were close to each other when the BB framework was used (see Figures 19 and 23). For AP Biology 2005–2007, the IRT true- and observed-score equating methods appeared to satisfy FOE better than other traditional curvilinear equating methods.

Figures 20, 22, 24, and 26 illustrate the differences in conditional expected scores under the IRT framework. The IRT true- and observed-score equating methods outperformed the other equating methods in terms of FOE when the IRT framework was used. For AP Biology 2004–2006, the traditional curvilinear equating methods seemed to perform better than the

linear equating methods in preserving FOE. For AP Biology 2005–2007, the Tucker equating method also appeared to satisfy FOE throughout the whole ability level.

Table 7 reports how well each equating method satisfied FOE under the BB and IRT frameworks using the overall FOE index (D_1). Regardless of the weights used, Levine true- and observed-score equating methods produced lower D_1 values than any other equating methods when the BB framework was used. This finding is consistent with the theoretical results reported in Brennan (2010), which were summarized earlier in this paper. When the IRT framework was used, the IRT true- and observed-score equating methods outperformed other equating methods in terms of FOE, regardless of weighting. These finding are consistent with previous research findings (Brennan, 2010; Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005). Other than that, there were no big differences in the D_1 values among the other equating methods. Even though the differences in the D_1 values were small, three chained equipercentile methods (unsmoothed, presmoothed, and postmoothed) yielded smaller D_1 values than three corresponding frequency estimation methods, with few exceptions (the AP Biology 2004–2006 linkage with normal weights under the BB framework for both raw and scale scores).

Figures 27 to 30 illustrate the differences in CSEMs for the raw scores between two forms. The differences in CSEMs for the scale scores are also presented in Figures 31 to 34. When the BB framework was used (see Figures 27, 29, 31, and 33), regardless of score types, the three linear equating methods seemed to perform well in terms of SOE, even though this pattern is less apparent for AP Biology 2004–2006 when the raw scores were used. This finding is consistent with Brennan's (2010) finding that SOE is more likely to be satisfied under the linear equating than the curvilinear equating. For AP Biology 2005–2007, with regard to CSEMs for the raw scores, all of the equating methods appeared to satisfy SOE well, except for the unsmoothed frequency estimation and unsmoothed chained equipercentile methods in the lower part of the ability range (see Figure 29).

Figures 28, 30, 32, and 34 present the differences in CSEMs when the IRT framework was used. For both of the linkages, when the raw scores were used, the three linear equating methods and IRT true- and observed-score equating methods performed well in terms of SOE (see Figure 28 and 30). For AP Biology 2004–2006, when the scale scores were used, all the equating methods seemed to preserve SOE well in the middle ability range. In the two ends of the ability, however, it is difficult to tell which method preserved SOE better (see Figure 32). For AP Biology 2005–2007, in regard to CSEMs of the scale scores, the IRT true- and observed- score equating methods and the Tucker method appeared to satisfy SOE better than the other equating methods.

To further compare how well SOE holds, we computed the overall SOE index (D_2) and presented the values in Table 8. Regardless of weighting and the psychometric framework used, the D_2 values for three linear equating methods were smaller than those for the other equating methods when raw scores were used. However, the differences in the D_2 values among different equating methods were relatively small. When the scale scores were used, it is difficult to conclude which method produces smaller D_2 values, and there are also no substantial differences in the D_2 values for all the equating methods.

Relationship Between Reliability and Equity Properties

In analyzing the pseudo-tests for the three reliability conditions, both the BB and IRT frameworks were used, and the results compared.

Figures 35 to 37 show the differences in conditional expected composite raw scores for the three reliability conditions under the BB framework. The differences in conditional expected scale scores for the three reliability conditions under the BB framework are presented in Figures 41 to 43. Regardless of the score types, the Levine true- and observed-score equating methods preserved FOE for all three reliability conditions, as expected from previous research (Brennan, 2010). The Tucker method seemed to preserve FOE better in the high- and medium-reliability conditions than in the low-reliability condition. It should be noted that according to Brennan (2010), "high-reliability" is not required to achieve FOE when the true equating function is linear, FOE holds for the applied true-score equating, and "equal reliabilities" guarantees that FOE is satisfied for the linear observed-score equating. In this study, alpha coefficients for the two forms were quite similar for all three reliability conditions. Thus, our result was consistent with Brennan's (2010). The curves for the curvilinear equating methods were close to each other in these graphs. It appears that the curves for the curvilinear equating methods were closer to the horizontal zero line in the high-reliability condition than in the other two reliability conditions. This finding is also consistent with Brennan (2010), who showed that "high-reliability" facilitates achieving approximate FOE when equating is curvilinear.

Figures 38 to 40 show the differences in conditional expected composite raw scores for the three reliability conditions under the IRT framework. The differences in conditional expected scale scores for the three reliability conditions under the IRT framework are presented in Figures 44 to 46. In the high-reliability condition, regardless of the score types, all the equating methods seemed to preserve FOE quite well, except the three linear equating methods in the lower ability range. In the medium- and low-reliability conditions, however, the IRT true score equating method outperformed other equating methods in terms of FOE.

The overall FOE index (D_1) values for all the reliability conditions are presented in Table 9. Under the IRT framework, regardless of weights used, the D_1 values for all the traditional curvilinear equating methods become bigger as the reliability decreases. When the scale scores were used under the BB framework, regardless of weights, the D_1 values for all the traditional curvilinear equating methods tended to be smaller in the high-reliability condition than in the low-reliability condition. This pattern for the traditional curvilinear equating methods was also found when the raw scores were used under the BB framework, with the exceptions of the three chained equipercentile methods when the normal weights were used. With respect to the linear equating methods, FOE was satisfied under the BB framework regardless of the magnitude of reliability. These findings are consistent with the previous research findings (Brennan, 2010). However, under the IRT framework, the linear equating methods performed worse than any other equating methods in terms of FOE, which may be due to the inconsistency in the model (which assumes a curvilinear relationship between true scores) and the equating methods (which are linear).

Figures 47 to 49 show the differences in CSEMs of the raw scores for the three reliability conditions under the BB framework. Figures 50 to 52 show these differences under the IRT framework. The differences in CSEMs of the scale scores for the three reliability conditions under the BB framework are presented in Figures 53 to 55. These differences under the IRT framework are presented in Figures 56 to 58. Regardless of the psychometric framework used, when the scale scores are used, the curves for all the equating methods seemed to be

closer to the horizontal zero line in the high-reliability condition than in the medium- and low-reliability condition (see Figures 53 to 55 and 56 to 58). However, this pattern was not clearly shown when the raw scores were used (see Figures 47 to 49 and 50 to 52).

To further compare how well SOE holds in the three reliability conditions, we computed the overall SOE index (D_2) values and presented the values in Table 10. With regard to the CSEMs of the scale scores, under the IRT framework, the D_2 values for all the equating methods were smaller in the high-reliability condition than in the low-reliability condition, with the exception of the Tucker method with uniform weights and the unsmoothed frequency estimation method with normal weights. With respect to the CSEMs of the raw scores, under the IRT framework, the D_2 values for two IRT equating methods were smaller in the high-reliability condition than the low-reliability condition. Otherwise, there were no special findings regarding the reliability conditions.

Discussion

Equity Properties for Mixed-Format Tests

The current analyses for the intact forms with an IRT framework of AP Exams show that the IRT true-score equating tends to preserve FOE better than any other equating methods. With respect to SOE, the IRT observed-score equating tends to preserve SOE better than the IRT true-score equating method. These findings are consistent with previous research evidence (Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005), although these previous studies were for a different equating design (i.e., a random-groups design) and the tests used for those studies were not mixed-format tests. However, the values of the overall FOE indices for the equating methods that can be grouped together[5] (e.g., the IRT true- and observed- score equating methods can be grouped together because both of them are based on IRT) did not differ much from each other. The two IRT equating methods performed better than all the other equating methods in terms of FOE. This is because the equating methods and the evaluation framework were based on the same IRT model. In terms of SOE, there were also no big differences among the values of the overall SOE indices for various equating methods, and it did not seem that one method performed better than the others.

Effects of Psychometric Models on Equity Properties

When we used only the MC items of AP Exams to compare the results based on different psychometric frameworks, the three linear equating methods preserved FOE better than any other equating methods under the BB framework, and the two IRT equating methods outperformed other equating methods in terms of FOE under the IRT framework. These results are consistent with previous research findings (Brennan, 2010; Kim et al., 2005; Lee et al., 2010; Tong & Kolen, 2005). In this study, FOE was satisfied not only for the Levine true-score equating method but also for the Levine observed-score equating method under the BB framework, because the reliabilities for the two forms were quite similar (for AP Biology 2004–2006, coefficient α for the new form was .942 and for the old form was .945, and for AP Biology 2005–2007, coefficient α for the new form was .945 and for the old form was .936; see Table 3). If the reliabilities for the two forms are equal, FOE is satisfied for the observed-score linear equating as well as for the true-score linear equating (Brennan, 2010).

5. The equating methods considered in this study can be grouped into three categories: linear equating methods, traditional curvilinear equating methods, and IRT equating methods.

With regard to SOE, the D_2 values for the three linear equating methods were smaller than those for the other equating methods regardless of the psychometric framework used in the raw-score scale. Comparing the two IRT equating methods, IRT observed-score equating performed better than IRT true-score equating in terms of SOE, which is consistent with the previously discussed results for the mixed-format intact forms. To summarize, we used the IRT and BB frameworks to examine whether different underlying psychometric models would yield different results for expected scores and CSEMs. Our findings suggest that the performance of various equating methods in terms of FOE and SOE depends on the psychometric model assumed, which echoes Kim et al.'s (2005) findings.

Relationship Between Reliability and Equity Properties

Analyses of the shortened MC pseudo-forms show that the three linear equating methods preserve FOE well, regardless of the reliability conditions and score types, when the BB framework is used. This finding is consistent with Brennan (2010), who showed that FOE and SOE do not directly depend upon the magnitude of reliability for the linear equating, and FOE and SOE are more likely satisfied under the linear equating than the curvilinear equating. When the IRT framework is used, in the high-reliability condition, all equating methods seem to preserve FOE quite well except the three linear equating methods in the lower ability area, regardless of the score types. In the medium- and low-reliability conditions, however, IRT true score equating outperforms other equating methods in terms of FOE. Also, regardless of the psychometric framework and weights used, the D_1 values for all the traditional curvilinear equating methods increase as the reliability decreases, with one exception: when the raw score scale is used under the BB framework. This finding is consistent with previous research, which showed that for curvilinear equating, FOE and SOE are not generally satisfied, but the magnitude of reliability matters in the sense that all other things being equal, FOE and SOE are more nearly satisfied when reliabilities are high (Brennan, 2010). With respect to SOE, when the scale scores are used, the curves for all the equating methods seem to be closer to the horizontal zero line in the high-reliability condition than in the medium- and low-reliability conditions, regardless of the psychometric framework used. Also, under the IRT model, the D_2 values for all the equating methods are smaller in the high-reliability condition than in the low-reliability condition, with two previously discussed exceptions when the scale scores are used.

Overall, our findings for FOE and its relationship with reliability are consistent with previous research (Brennan, 2010). On the other hand, the pattern of findings for SOE is less consistent than that for FOE. According to Brennan (2010), SOE is satisfied under certain conditions, as described earlier. Those conditions for SOE are not likely to be met in our study. For example, the assumption of homogeneous error variances seems unlikely to be met, not only in our study but also in most circumstances.

Conclusion

This study set out to examine the performance of various equating methods using AP Exam data. In particular, we assessed the equity properties of various equating methods using mixed-format tests with the common-item, nonequivalent-groups design; examined the effect of underlying psychometric models on the equity properties using both the IRT and BB frameworks; and investigated the relationships between test reliability and equity properties using pseudo-tests. Taken together, our findings contribute to understanding of the equity properties by providing a more comprehensive set of empirical results with different test formats, designs, and test forms.

Despite our contributions, our results should be interpreted and applied with caution. We used the IRT and BB frameworks and found that the equity properties of various equating methods are influenced by the underlying psychometric model assumed. Many practical reasons may make the use of one of the methods preferable to others in a particular context. It should be noted, however, that when the IRT framework is used, the IRT true-score equating might have potential advantages over the traditional equating methods in terms of FOE because the true-score distributions are based on the IRT model. Similarly, all the linear equating methods might have potential advantages over all the curvilinear equating methods when the BB framework is used because the true-score relationship is assumed to be linear. Furthermore, there are strong assumptions in using the IRT framework, and it is important to note that the results depend on how well the IRT model fits the data. In addition, as Lee et al. (2010) suggest, FOE and SOE should not be the only criteria to assess the equating results. As Brennan (2010) described, FOE and SOE are more likely to be satisfied under linear equating than curvilinear equating only when the true equating function is linear, which is seldom the case. Thus, even though the linear equating methods preserve FOE and SOE better than the curvilinear equating methods, it does not mean that the linear equating methods are preferable.

In this study, the true-score distribution is assumed to be a symmetric beta distribution under the BB framework and a normal distribution under the IRT framework. If a true-score distribution is not symmetric, then the results based on a symmetric true-score distribution may not be directly applicable. However, by considering both uniform weights and normal weights in computing the overall discrepancy indices (D_1 and D_2), the assumption of the symmetric true-score distribution is unlikely to alter our conclusions because no significant differences are observed for the two types of weights.

This paper examines how FOE and SOE hold in terms of both raw and scale scores. We also use two types of weights when we compute the values of overall FOE and SOE indices. There are no significantly different findings across the score types or weights. Due to the scope of this study, a limited number of factors are investigated. Future research could incorporate other factors that are not considered in the current study.

Like other studies, there are some limitations in this study. First of all, when we eliminated items to construct pseudo-tests for the three reliability conditions, we only considered item-total correlations. Because other test and item characteristics, such as a content specification or the correlations between common items and total scores, are not considered in constructing the pseudo-forms, the shortened forms may not be parallel, especially in terms of content representativeness. In other words, the shortened common-item sets may not be a "mini-version" of the total test (Kolen & Brennan, 2004). Therefore, it may be difficult to claim that strictly adequate equating results were obtained when the shortened pseudo-forms were analyzed. Second, all the items in the original AP Exams have high item-total correlations. So when we constructed pseudo-forms for the three reliability conditions, it was difficult to allow sufficient variability in reliabilities among those forms to examine the relationships between reliability and FOE and SOE. Furthermore, we defined coefficient α as reliability in constructing and analyzing pseudo-forms for the three reliability conditions. Use of different reliability statistics might affect the results. Future research with a simulation study seems to be necessary to explore the relationship between reliability and the preservation of equating properties.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorridike (Ed.), *Educational measurement*. 2nd ed. (pp. 508–600). Washington, DC: American Council on Education.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Brennan, R. L. (2010). *First-order and second-order equity in equating* (CASMA Research Report No. 30). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Han, T., Kolen, M. J., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equating and traditional equipercentile equating. *Applied Measurement in Education, 10*, 105–121.

Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics, 16*, 93–100.

Kim, D. I., Brennan, R. L., & Kolen, M. J. (2005). A comparison of IRT equating and beta 4 equating. *Journal of Educational Measurement, 42*(1), 77–99.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1–11.

Kolen, M. J. (2004). *POLYCSEM* (*Windows Console version*) [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Kolen, M. H., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 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., & 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, University of Iowa.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

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

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph, No. 17*. Richmond, VA: Psychometric Society.

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

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog* (*version 7*) [Computer software]. Lincolnwood, IL: Scientific Software International.

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

Wang, T., Hanson, B. A., & Harris, D. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24*, 195–210.

Appendix

Table A1.

Weights Assigned to MC and CR Items (Intact Forms)

Test	Form	MC Weights	CR Weights	Weighted MC	Weighted CR	Composite	% MC	% CR
Biology	2004	2	3	198	120	318	0.62	0.38
	2006	2	3	196	120	316	0.62	0.38
	2005	2	3	196	120	316	0.62	0.38
	2007	2	3	198	120	318	0.63	0.38
English Language and Composition	2004	2	5	106	135	241	0.44	0.56
	2007	2	5	104	135	239	0.44	0.56
French Language and Culture	2005	2	1,1,5,3,3,3,3,3	158	150	308	0.51	0.49
	2007	2	1,1,5,3,3,3,3,3	170	150	320	0.53	0.47

Table A2.

Descriptive Statistics for Weighted Composite and CI Scores (Intact Form)

Test	Form		N	Mean	Std Dev	Skewness	Kurtosis	Corr b/w Composite & CI
Biology	2004	Composite	15,075	178.002	57.491	-0.274	-0.632	0.908
		CI	15,075	33.079	10.439	-0.470	-0.478	
	2006	Composite	16,899	179.056	57.409	-0.375	-0.569	0.910
		CI	16,899	32.127	10.210	-0.408	-0.464	
	2005	Composite	16,185	176.451	53.387	-0.320	-0.466	0.873
		CI	16,185	30.450	8.608	-0.573	-0.236	
	2007	Composite	16,819	164.540	56.069	-0.102	-0.693	0.887
		CI	16,819	29.905	9.222	-0.493	-0.458	
English Language and Composition	2004	Composite	15,820	147.768	33.850	-0.460	-0.003	0.757
		CI	15,820	27.700	6.295	-1.075	0.988	
	2007	Composite	16,882	142.077	33.110	-0.427	0.092	0.765
		CI	16,882	27.648	6.179	-1.035	0.959	
French Language and Culture	2005	Composite	13,571	182.708	50.157	-0.321	-0.445	0.900
		CI	13,571	32.309	10.327	-0.320	-0.727	
	2007	Composite	13,982	182.125	51.325	-0.260	-0.443	0.893
		CI	13,982	31.751	10.182	-0.240	-0.745	

Table A3.

Descriptive Statistics for Unweighted MC and CI Scores (Full-Length MC Pseudo-Tests)

Test	Form		# of items	N	Mean	Std Dev	Skew-ness	Kurtosis	α coeff.	Corr b/w Total & CI
Biology	2004	MC Total	99	15,075	61.669	17.662	-0.398	-0.584	0.945	0.935
		CI	26	15,075	16.539	5.219	-0.470	-0.478	0.834	
	2006	MC Total	98	16,899	62.156	16.759	-0.521	-0.315	0.942	0.934
		CI	26	16,899	16.064	5.105	-0.408	-0.464	0.830	
	2005	MC Total	98	16,185	65.235	15.985	-0.647	-0.098	0.936	0.909
		CI	23	16,185	15.225	4.304	-0.573	-0.236	0.778	
	2007	MC Total	99	16,819	63.856	17.371	-0.441	-0.487	0.945	0.920
		CI	23	16,819	14.952	4.611	-0.493	-0.458	0.807	

Table A4.

Descriptive Statistics for Unweighted MC and CI Scores (Shortened MC Pseudo-Tests)

Biology	Form	Variable	# of items	N	Mean	Std Dev	Skew-ness	Kurtosis	α coeff.	Corr b/w Total & CI
High Reliability	2004	MC	50	15,075	32.682	10.978	-0.496	-0.730	0.931	0.913
		CI	13	15,075	8.397	3.205	-0.515	-0.625	0.775	
	2006	MC	49	16,899	32.912	10.095	-0.654	-0.389	0.923	0.912
		CI	13	16,899	8.463	3.113	-0.534	-0.532	0.775	
Medium Reliability	2004	MC	50	15,075	27.892	9.129	-0.089	-0.762	0.888	0.881
		CI	13	15,075	7.271	2.823	-0.117	-0.697	0.693	
	2006	MC	49	16,899	29.806	8.542	-0.365	-0.481	0.884	0.878
		CI	13	16,899	7.180	2.805	-0.095	-0.666	0.688	
Low Reliability	2004	MC	50	15,075	29.758	7.445	-0.238	-0.162	0.839	0.854
		CI	13	15,075	8.143	2.463	-0.380	-0.193	0.631	
	2006	MC	49	16,899	29.244	7.263	-0.278	-0.065	0.839	0.851
		CI	13	16,899	7.601	2.441	-0.224	-0.256	0.627	

Table A5.

Overall FOE Index (D_j) for Various Equating Methods (IRT Framework, Intact Forms)

	Weights	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores												
Biology: 2004–2006	Uniform	2.53	2.78	2.83	2.44	2.21	2.61	2.40	2.34	2.15	***0.12***	0.50
	Normal	1.59	1.09	1.16	1.46	1.39	1.50	1.38	1.44	1.37	***0.09***	0.24
Biology: 2005–2007	Uniform	5.31	3.70	3.89	5.16	4.77	5.30	4.88	5.36	4.96	***0.19***	0.39
	Normal	4.62	4.79	4.82	3.83	3.68	3.87	3.67	3.84	3.70	***0.08***	0.19
English Language and Composition	Uniform	1.22	1.20	1.20	0.88	0.95	0.85	0.93	0.86	0.94	***0.14***	0.61
	Normal	0.59	0.77	0.54	0.36	0.37	0.37	0.38	0.36	0.38	***0.20***	0.28
French Language and Culture	Uniform	9.69	9.07	9.04	9.46	9.85	9.97	10.10	9.16	9.33	***0.14***	0.18
	Normal	9.58	8.93	8.89	9.58	9.39	9.64	9.42	9.51	9.31	***0.08***	0.15
Scale Scores												
Biology: 2004–2006	Uniform	0.87	1.05	1.03	0.86	0.76	0.91	0.82	0.82	0.74	***0.04***	0.12
	Normal	0.30	0.25	0.25	0.38	0.34	0.39	0.35	0.37	0.33	***0.01***	0.04
Biology: 2005–2007	Uniform	1.54	0.99	1.06	1.67	1.53	1.71	1.58	1.75	1.61	***0.06***	0.12
	Normal	1.02	0.97	0.99	0.94	0.88	0.95	0.88	0.94	0.89	***0.02***	0.04
English Language and Composition	Uniform	0.50	0.42	0.49	0.36	0.36	0.36	0.36	0.35	0.36	***0.05***	0.22
	Normal	0.20	0.23	0.18	0.12	0.13	0.13	0.12	0.12	0.13	***0.08***	0.09
French Language and Culture	Uniform	2.71	2.54	2.51	2.76	2.81	2.83	2.89	2.59	2.66	***0.05***	0.06
	Normal	2.19	2.03	2.00	2.20	2.17	2.22	2.19	2.16	2.13	***0.02***	0.03

Note*: Tucker = Tucker method, Lev Ob = Levine observed-score method, Lev Tr = Levine true-score method, FE = unsmoothed frequency estimation method, ChainedE = unsmoothed chained equipercentile method, Pre Fe = frequency estimation method with log-linear presmoothing, Pre Ch = chained equipercentile method with log-linear presmoothing, Post Fe = frequency estimation method with cubic-spline postsmoothing, Post Ch = chained equipercentile method with cubic-spline postsmoothing, IRT Tr = IRT true-score method, IRT Ob = IRT observed-score method; figures in italic and bold indicate the corresponding equating method yields the smallest discrepancy for a test.

* This note is applicable to Tables 5 through 10.

Table A6.

Overall SOE Index (D_2) for Various Equating Methods (IRT Framework, Intact Forms)

	Weights	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores												
Biology: 2004–2006	Uniform	5.76	5.96	5.78	5.25	5.17	5.22	***5.17***	5.38	5.33	5.52	5.36
	Normal	5.96	6.20	5.98	***5.20***	5.44	5.21	5.53	5.25	5.47	6.04	5.89
Biology: 2005–2007	Uniform	5.23	5.35	5.32	4.23	3.99	4.00	***3.36***	3.85	3.68	4.41	4.32
	Normal	4.29	4.59	4.52	3.50	3.11	3.48	***2.36***	3.44	3.00	4.84	4.71
English Language amd Composition	Uniform	2.39	3.64	***2.29***	3.79	3.82	3.50	3.55	3.94	4.01	4.28	3.48
	Normal	2.35	3.71	***2.22***	2.50	2.80	2.62	2.88	2.52	2.83	3.48	2.71
French Language and Culture	Uniform	2.90	2.74	2.67	3.70	3.62	2.97	2.80	3.72	3.72	2.69	***2.56***
	Normal	2.07	1.85	***1.73***	2.50	2.66	2.11	1.81	2.60	2.61	3.12	3.04
Scale Scores												
Biology: 2004–2006	Uniform	1.67	1.77	1.74	1.57	1.50	1.49	1.42	1.49	1.46	1.36	***1.26***
	Normal	1.24	1.32	1.27	0.97	1.01	1.00	1.04	***0.93***	0.98	1.14	1.10
Biology: 2005–2007	Uniform	1.61	1.53	1.54	1.32	1.26	1.25	1.15	1.21	***1.13***	1.27	1.23
	Normal	0.84	0.95	0.93	0.93	0.79	0.93	***0.75***	0.93	0.79	1.16	1.11
English Language and Composition	Uniform	***0.87***	1.00	0.87	1.08	1.07	0.94	0.96	1.17	1.17	1.57	1.22
	Normal	0.72	1.08	***0.71***	0.78	0.85	0.80	0.87	0.79	0.86	1.25	0.96
French Language and Culture	Uniform	1.43	1.36	1.34	1.54	1.56	1.56	1.63	1.35	1.39	0.73	***0.66***
	Normal	0.91	0.83	0.80	0.99	1.03	1.02	1.07	0.89	0.94	0.71	***0.68***

Table A7.

Overall FOE Index (D_1) for Various Equating Methods (Full-Length MC Pseudo-Tests)

	Framework	Weights	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores													
Biology: 2004–2006	BB	Uniform	0.17	0.03	***0.00***	1.17	1.11	1.17	1.12	1.19	1.16	0.99	0.80
		Normal	0.17	0.02	***0.00***	0.56	0.60	0.57	0.61	0.56	0.61	0.53	0.50
	IRT	Uniform	1.07	1.15	1.14	0.30	0.23	0.30	0.25	0.32	0.25	***0.04***	0.17
		Normal	0.60	0.54	0.53	0.29	0.23	0.30	0.24	0.29	0.24	***0.04***	0.05
Biology: 2005–2007	BB	Uniform	0.62	0.02	***0.00***	0.69	0.57	0.56	0.52	0.62	0.61	0.31	0.29
		Normal	0.55	0.01	***0.00***	0.65	0.52	0.61	0.50	0.65	0.53	0.30	0.29
	IRT	Uniform	0.62	0.45	0.45	0.65	0.51	0.62	0.47	0.62	0.48	***0.03***	0.05
		Normal	0.43	0.35	0.36	0.54	0.35	0.55	0.37	0.53	0.35	***0.03***	0.05
Scale Scores													
Biology: 2004–2006	BB	Uniform	0.11	0.03	***0.03***	1.03	0.96	1.04	0.98	1.06	1.02	0.90	0.70
		Normal	0.08	***0.01***	***0.01***	0.33	0.35	0.33	0.35	0.33	0.35	0.30	0.27
	IRT	Uniform	1.24	1.42	1.41	0.31	0.25	0.34	0.27	0.35	0.27	***0.08***	0.16
		Normal	0.40	0.41	0.40	0.24	0.19	0.26	0.20	0.24	0.19	0.04	***0.03***
Biology: 2005–2007	BB	Uniform	0.43	0.02	***0.01***	0.38	0.35	0.40	0.38	0.44	0.46	0.18	0.20
		Normal	0.29	0.01	***0.00***	0.34	0.28	0.32	0.28	0.34	0.30	0.16	0.16
	IRT	Uniform	0.48	0.54	0.54	0.76	0.61	0.71	0.55	0.69	0.53	***0.02***	0.04
		Normal	0.26	0.27	0.27	0.39	0.27	0.42	0.28	0.39	0.26	***0.02***	0.03

Table A8.

Overall SOE Index (D_2) for Various Equating Methods (Full-Length MC Pseudo-Tests)

	Framework	Weights	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores													
Biology: 2004–2006	BB	Uniform	1.26	***1.20***	1.21	1.64	1.57	1.60	1.52	1.62	1.56	1.52	1.28
		Normal	1.16	***1.04***	1.07	1.41	1.35	1.43	1.39	1.41	1.35	1.30	1.20
	IRT	Uniform	***0.66***	0.67	0.67	1.53	1.50	1.52	1.49	1.54	1.54	1.51	1.34
		Normal	***0.63***	0.72	0.69	1.39	1.41	1.43	1.42	1.40	1.42	1.18	1.10
Biology: 2005–2007	BB	Uniform	0.87	0.58	***0.57***	1.99	1.71	1.37	1.34	1.39	1.43	1.19	1.05
		Normal	0.96	0.49	***0.46***	1.46	1.57	1.45	1.45	1.42	1.53	1.30	1.25
	IRT	Uniform	0.83	***0.64***	0.64	1.20	1.23	1.08	1.03	1.21	1.22	0.86	0.76
		Normal	0.83	0.70	***0.70***	1.27	1.26	1.23	1.10	1.26	1.23	0.89	0.84
Scale Scores													
Biology: 2004–2006	BB	Uniform	***0.91***	0.92	0.93	1.52	1.49	1.47	1.41	1.50	1.46	1.40	1.20
		Normal	0.65	***0.59***	0.61	0.92	0.89	0.92	0.89	0.92	0.89	0.84	0.74
	IRT	Uniform	1.22	1.31	1.31	1.24	1.15	1.29	1.24	1.32	1.32	1.40	***1.16***
		Normal	***0.63***	0.74	0.73	0.99	0.98	1.04	1.01	1.02	1.01	0.88	0.78
Biology: 2005–2007	BB	Uniform	***2.90***	2.99	2.99	2.88	2.87	2.89	2.92	2.92	2.95	2.92	2.95
		Normal	3.15	3.27	3.26	3.10	***3.06***	3.14	3.15	3.17	3.20	3.15	3.20
	IRT	Uniform	0.87	0.98	0.97	1.20	1.19	1.09	1.03	1.22	1.15	0.73	***0.62***
		Normal	0.59	0.71	0.71	1.04	0.97	1.03	0.92	1.02	0.94	0.59	***0.55***

Table A9.

Overall FOE Index (D_j) for Various Equating Methods (Shortened MC Pseudo-Tests)

	Weights	Rel.	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores													
BB	Uniform	H	0.09	0.06	***0.00***	0.49	0.47	0.47	0.46	0.48	0.46	0.42	0.35
		M	0.07	0.02	***0.00***	0.80	0.79	0.82	0.81	0.80	0.80	0.63	0.48
		L	0.51	0.00	***0.00***	0.81	0.66	0.81	0.66	0.80	0.65	0.40	0.21
	Normal	H	0.08	0.04	***0.00***	0.42	0.37	0.43	0.39	0.42	0.37	0.32	0.30
		M	0.06	0.01	***0.00***	0.43	0.45	0.47	0.49	0.43	0.45	0.34	0.32
		L	0.51	0.00	***0.00***	0.50	0.30	0.49	0.29	0.49	0.29	0.13	0.07
IRT	Uniform	H	0.52	0.52	0.51	0.13	0.15	0.15	0.17	0.13	0.15	***0.04***	0.11
		M	0.86	0.88	0.88	0.23	0.22	0.24	0.23	0.25	0.23	***0.06***	0.21
		L	0.50	0.46	0.46	0.46	0.33	0.48	0.35	0.45	0.32	***0.07***	0.15
	Normal	H	0.26	0.31	0.29	0.12	0.14	0.13	0.13	0.12	0.14	***0.03***	0.05
		M	0.44	0.41	0.41	0.20	0.19	0.23	0.23	0.20	0.20	***0.06***	0.07
		L	0.54	0.18	0.18	0.53	0.30	0.54	0.30	0.53	0.30	***0.06***	0.08
Scale Scores													
BB	Uniform	H	0.13	0.04	***0.04***	0.63	0.62	0.59	0.58	0.62	0.61	0.59	0.46
		M	0.11	0.06	***0.06***	1.33	1.31	1.34	1.31	1.33	1.32	1.08	0.77
		L	0.72	0.04	***0.03***	1.28	1.08	1.25	1.04	1.25	1.04	0.68	***0.35***
	Normal	H	0.08	0.02	***0.02***	0.32	0.29	0.32	0.30	0.32	0.29	0.25	0.22
		M	0.07	0.03	***0.02***	0.49	0.50	0.52	0.54	0.49	0.50	0.38	0.34
		L	0.65	0.01	***0.01***	0.66	0.40	0.65	0.39	0.66	0.39	0.18	0.10
IRT	Uniform	H	0.92	0.94	0.90	0.29	0.31	0.31	0.32	0.31	0.32	***0.09***	0.17
		M	1.62	1.68	1.69	0.45	0.41	0.43	0.39	0.50	0.47	***0.17***	0.37
		L	0.88	0.99	0.99	0.71	0.54	0.74	0.56	0.69	0.51	***0.14***	0.27
	Normal	H	0.31	0.36	0.32	0.18	0.19	0.18	0.17	0.18	0.19	***0.05***	0.05
		M	0.55	0.53	0.53	0.30	0.29	0.33	0.32	0.30	0.30	0.11	***0.09***
		L	0.74	0.28	0.28	0.74	0.42	0.76	0.44	0.74	0.43	***0.10***	0.12

Note: H = high-reliability condition, M = medium-reliability condition, L = low-reliability condition.

Table A10.

Overall SOE Index (D_2) for Various Equating Methods (Shortened MC Pseudo-Tests)

	Weights	Rel.	Tucker	Lev Ob	Lev Tr	FE	ChainedE	Pre Fe	Pre Ch	Post Fe	Post Ch	IRT Tr	IRT Ob
Raw Scores													
BB	Uniform	H	1.11	***1.00***	1.05	1.18	1.10	1.12	1.07	1.16	1.09	1.15	1.03
		M	1.23	1.21	1.20	1.32	1.31	1.36	1.34	1.33	1.32	1.20	***0.94***
		L	0.90	0.89	0.89	0.82	0.93	0.77	0.89	0.77	0.87	0.98	***0.62***
	Normal	H	1.14	1.01	1.07	1.07	0.95	1.03	0.99	1.06	***0.94***	1.11	1.06
		M	1.10	1.07	1.05	1.28	1.27	1.39	1.39	1.28	1.28	1.09	***1.00***
		L	0.65	0.64	0.64	0.63	0.82	***0.60***	0.84	0.61	0.81	0.79	0.67
IRT	Uniform	H	0.66	***0.54***	0.60	0.92	0.87	0.92	0.88	0.93	0.89	0.97	0.85
		M	0.77	0.76	***0.76***	1.29	1.27	1.28	1.26	1.30	1.30	1.25	1.01
		L	***0.65***	0.69	0.69	0.93	1.06	0.89	1.00	0.93	1.05	1.09	0.88
	Normal	H	0.64	***0.48***	0.57	0.97	0.93	0.99	0.97	0.97	0.94	0.89	0.83
		M	0.64	***0.64***	0.64	1.25	1.26	1.31	1.32	1.26	1.27	1.01	0.88
		L	0.68	***0.74***	0.74	0.78	0.95	0.81	0.96	0.78	0.95	0.97	0.84
Scale Scores													
BB	Uniform	H	1.32	***1.10***	1.21	1.56	1.49	1.44	1.37	1.51	1.46	1.60	1.34
		M	1.80	1.80	1.78	2.16	2.10	2.12	2.06	2.18	2.15	2.10	***1.54***
		L	1.24	1.38	1.38	1.50	1.64	1.33	1.51	1.47	1.60	1.69	***1.14***
	Normal	H	0.98	***0.83***	0.90	0.95	0.89	0.91	0.88	0.94	0.88	0.98	0.89
		M	1.31	1.28	1.26	1.57	1.55	1.65	1.63	1.57	1.56	1.34	***1.15***
		L	***0.86***	0.89	0.89	0.96	1.19	0.90	1.17	0.94	1.17	1.15	0.92
IRT	Uniform	H	1.26	***1.06***	1.13	1.44	1.39	1.33	1.25	1.38	1.36	1.47	1.14
		M	1.81	1.83	1.83	2.09	2.01	2.05	1.97	2.17	2.14	2.18	***1.55***
		L	***1.23***	1.53	1.53	1.53	1.67	1.45	1.59	1.59	1.71	1.84	1.40
	Normal	H	0.78	***0.66***	0.67	1.19	1.18	1.15	1.12	1.18	1.18	1.00	0.89
		M	***1.03***	1.05	1.05	1.69	1.68	1.71	1.71	1.71	1.70	1.46	1.17
		L	***0.88***	1.12	1.11	1.17	1.38	1.20	1.40	1.18	1.39	1.45	1.20

Note: H = high-reliability condition, M = medium-reliability condition, L = low-reliability condition.

Figure A1.

Biology 2004–2006 unsmoothed and smoothed (S=0.1) equivalents using chained equipercentile method.

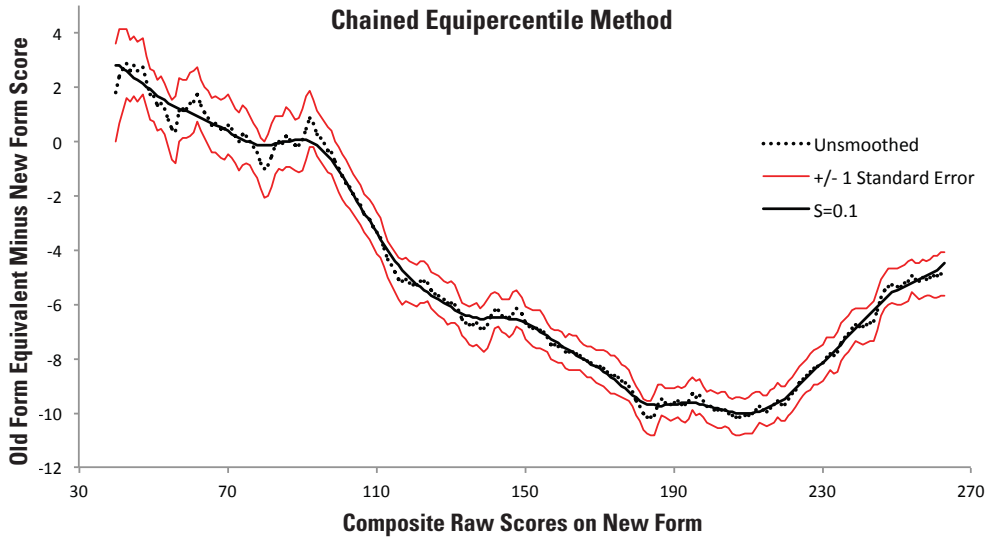


Figure A2.

Biology 2004–2006 unsmoothed and smoothed (S=0.1) equivalents using frequency estimation method.

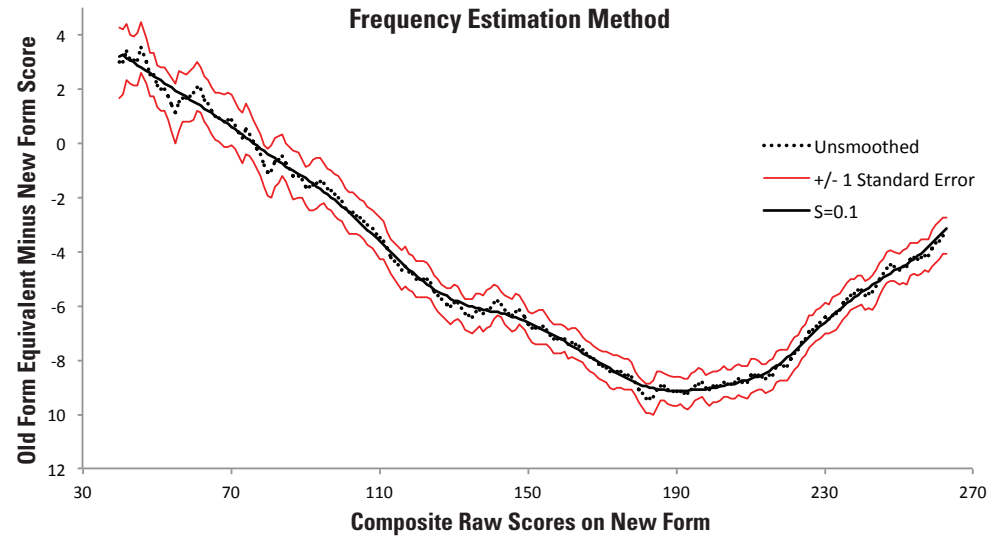


Figure A3.

Differences in expected composite raw scores for Biology 2004–2006 (IRT framework, intact forms).

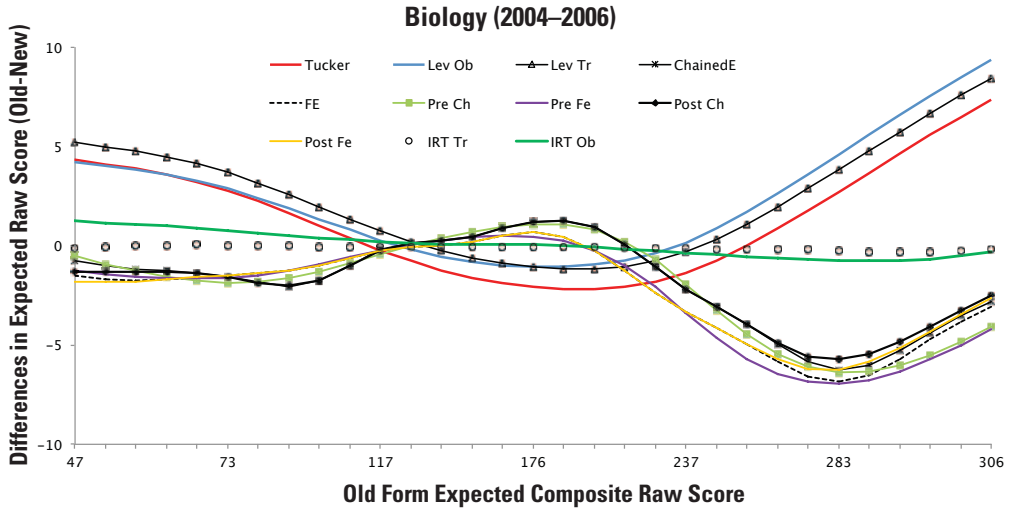


Figure A4.

Differences in expected composite raw scores for Biology 2005–2007 (IRT framework, intact forms).

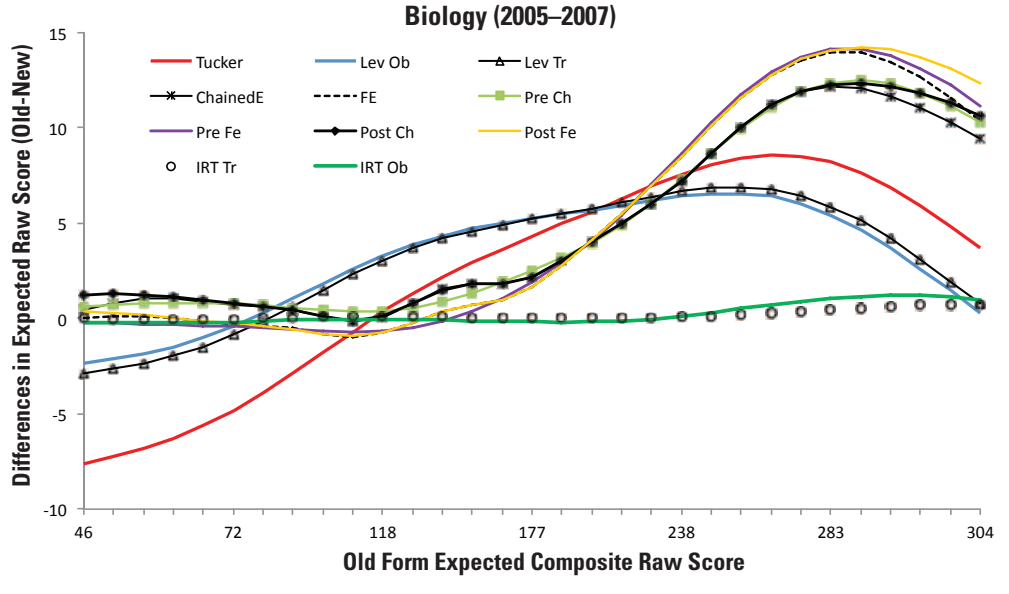


Figure A5.

Differences in expected composite raw scores for English Language and Composition 2004–2007 (IRT framework, intact forms).

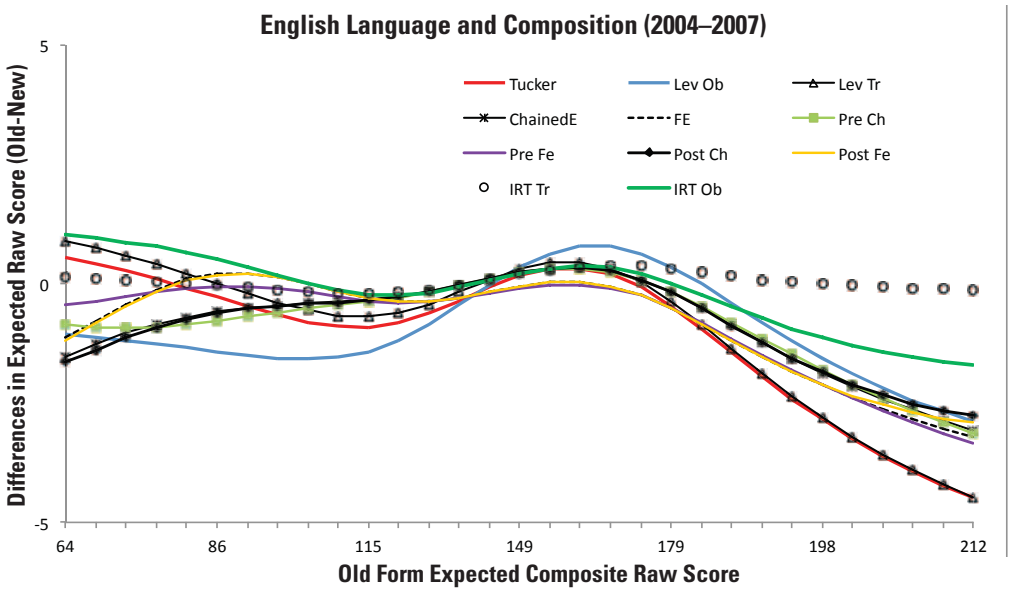


Figure A6.

Differences in expected composite raw scores for French Language and Culture 2005–2007 (IRT framework, intact forms).

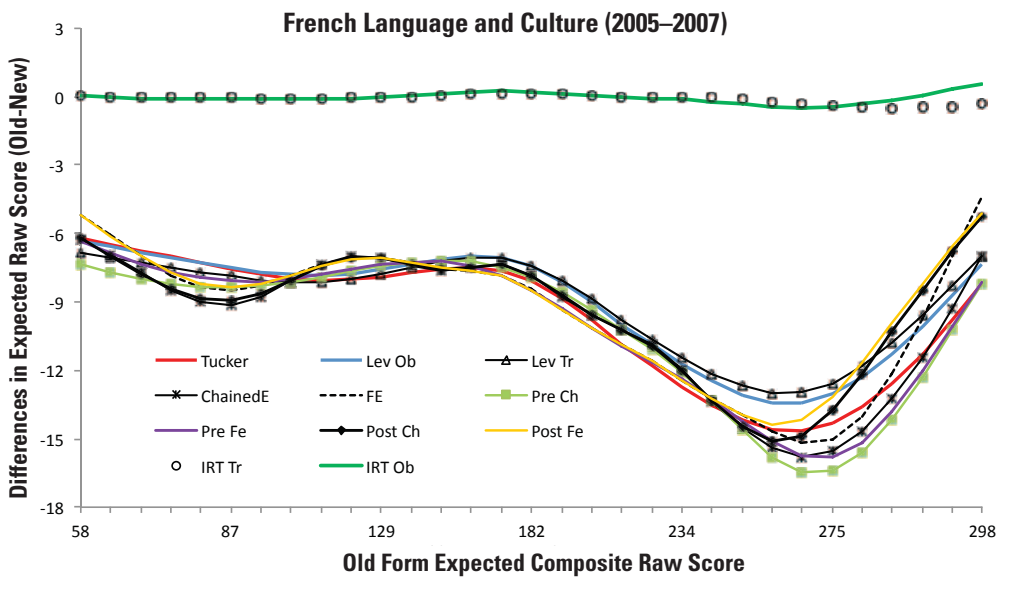


Figure A7.

Differences in expected composite scale scores for Biology 2004–2006 (IRT framework, intact forms).

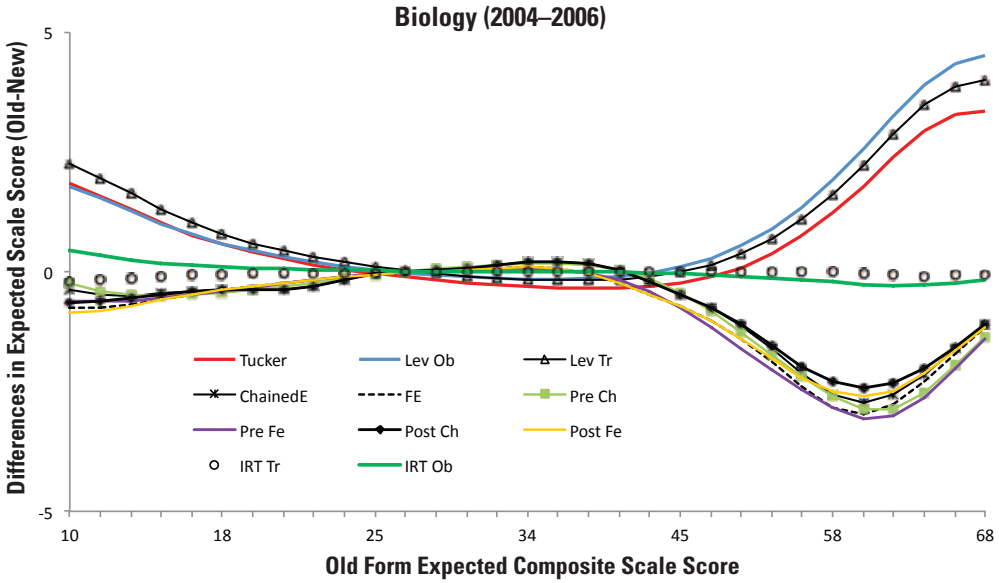


Figure A8.

Differences in expected composite scale scores for Biology 2005–2007 (IRT framework, intact forms).

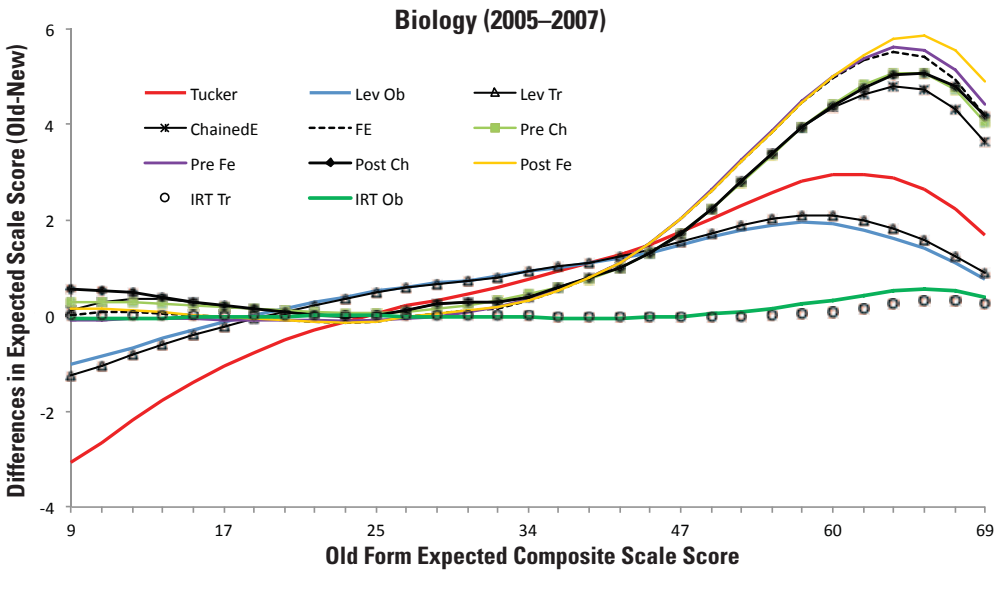


Figure A9.

Differences in expected composite scale scores for English Language and Composition 2004–2007 (IRT framework, intact forms).

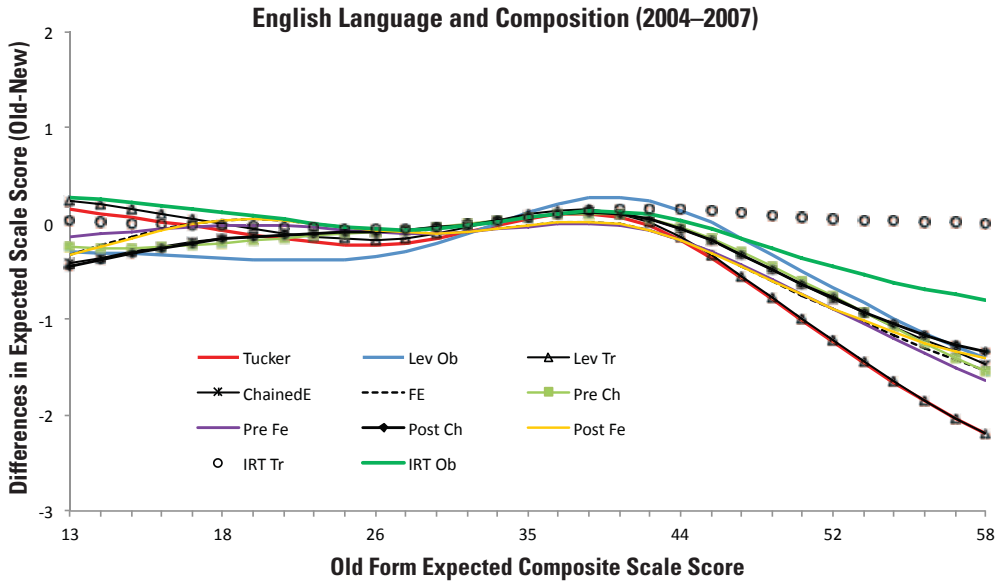


Figure A10.

Differences in expected composite scale scores for French Language and Culture 2005–2007 (IRT framework, intact forms).

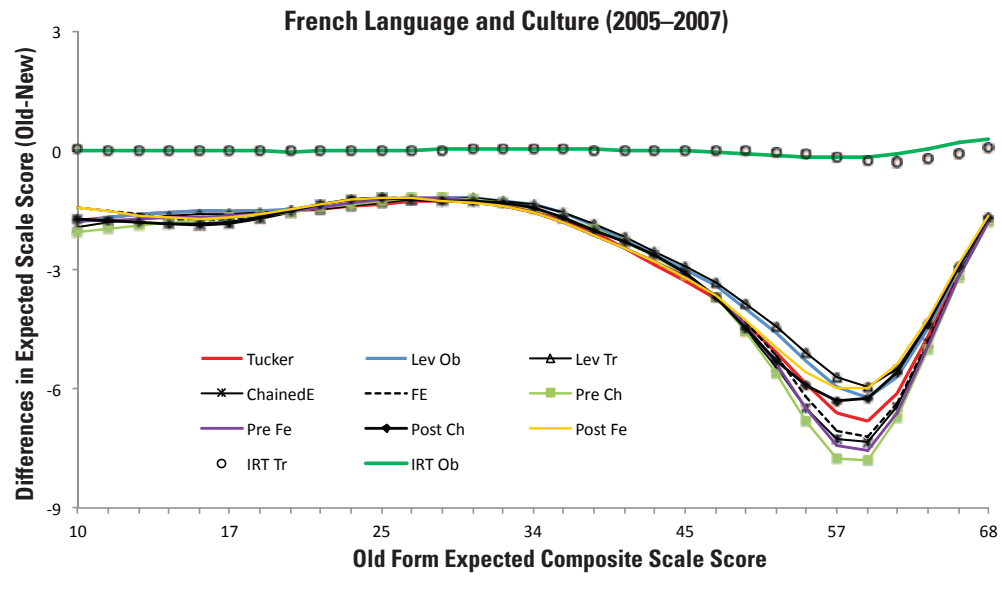


Figure A11.

Differences in CSEMs of composite raw scores for Biology 2004–2006 (IRT framework, intact forms).

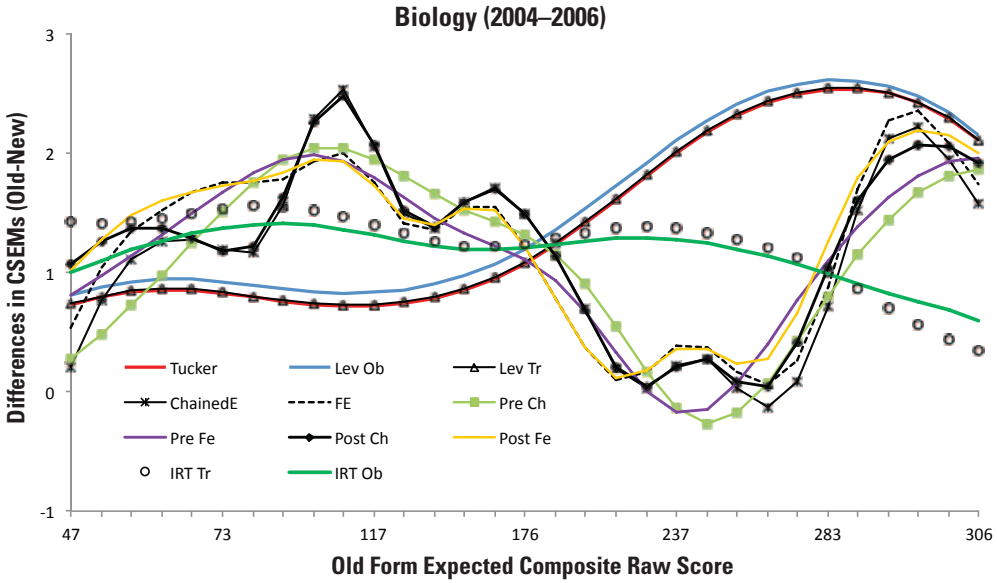


Figure A12.

Differences in CSEMs of composite raw scores for Biology 2005–2007 (IRT framework, intact forms).

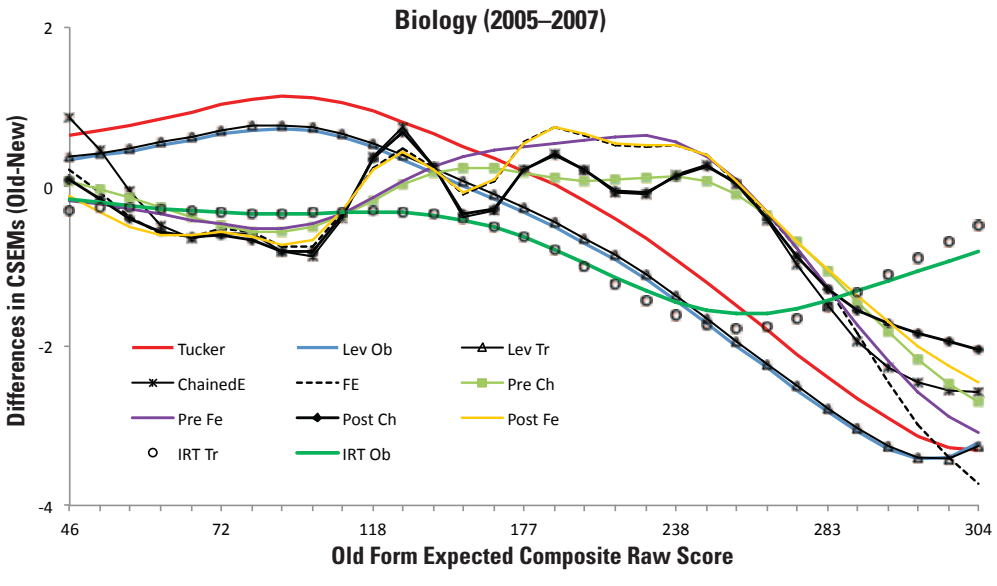


Figure A13.

Differences in CSEMs of composite raw scores for English Language and Composition 2004–2007 (IRT framework, intact forms).

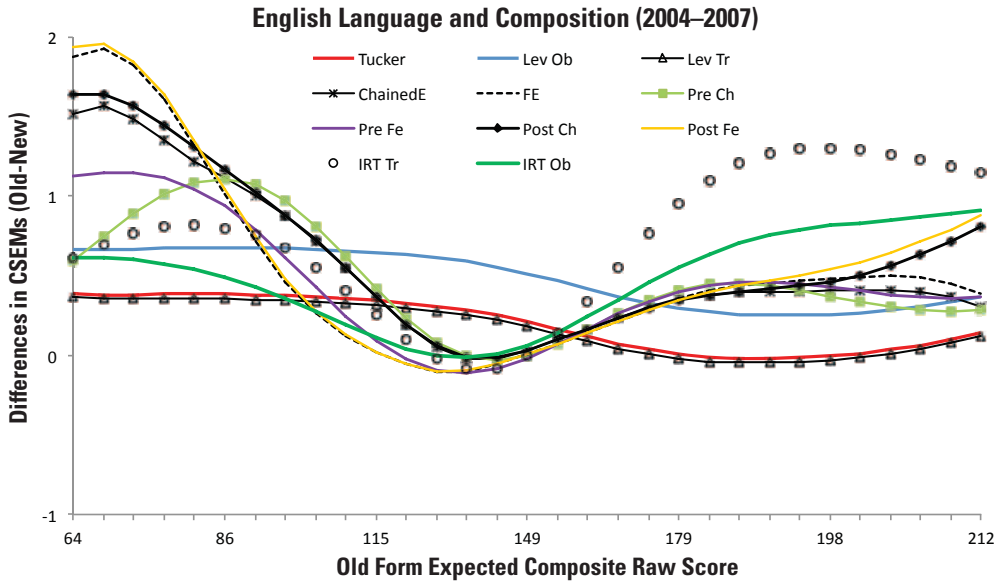


Figure A14.

Differences in CSEMs of composite raw scores for French Language and Culture 2005–2007 (IRT framework, intact forms).

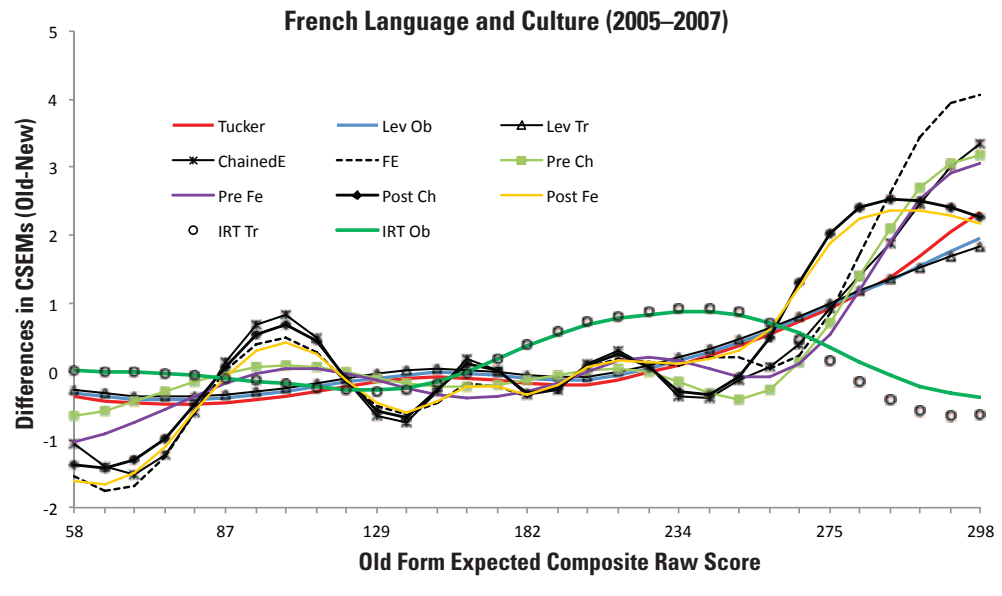


Figure A15.

Differences in CSEMs of composite scale scores for Biology 2004–2006 (IRT framework, intact forms).

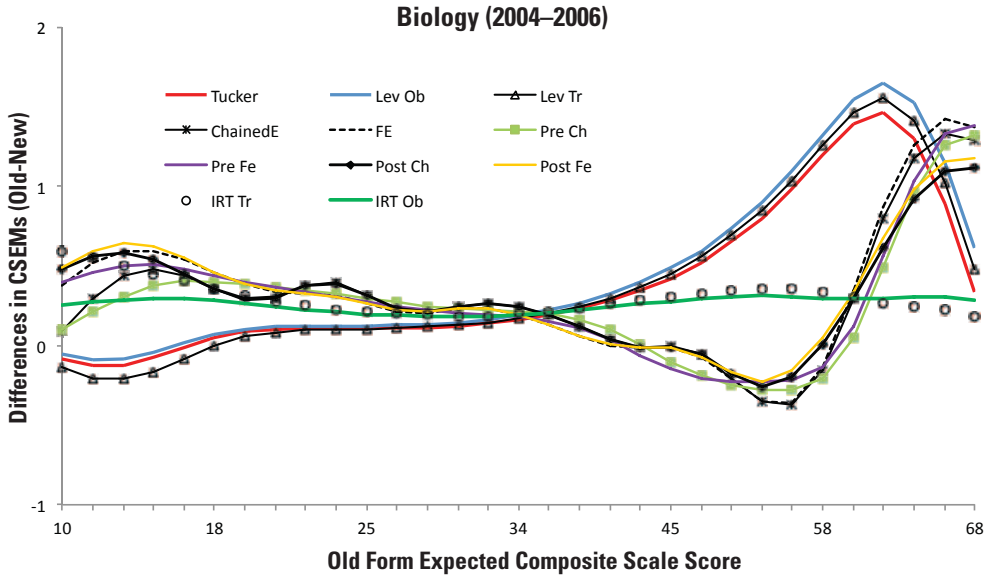


Figure A16.

Differences in CSEMs of composite scale scores for Biology 2005–2007 (IRT framework, intact forms).

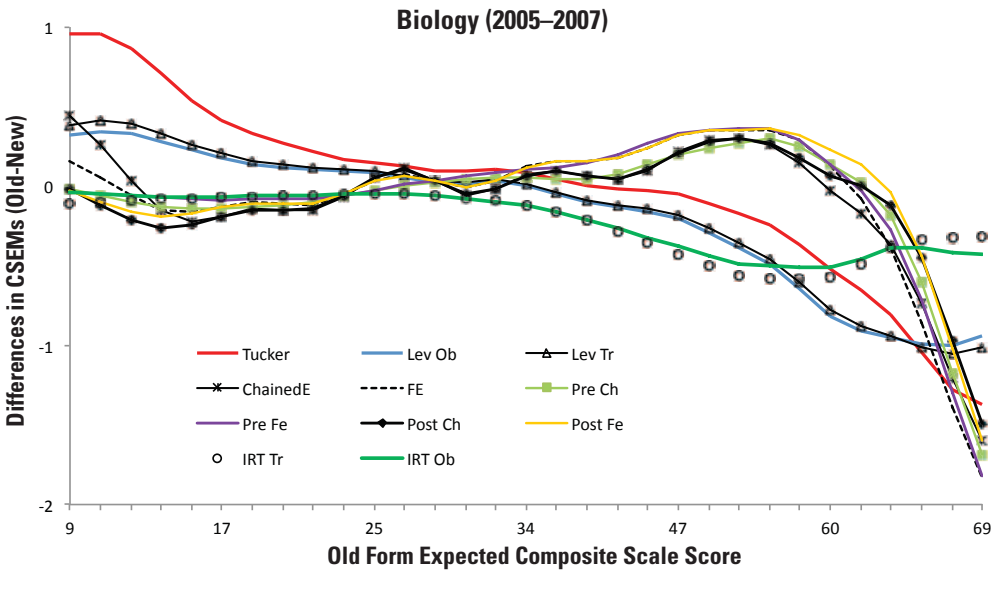


Figure A17.

Differences in CSEMs of composite scale scores for English Language and Composition 2004–2007 (IRT framework, intact forms).

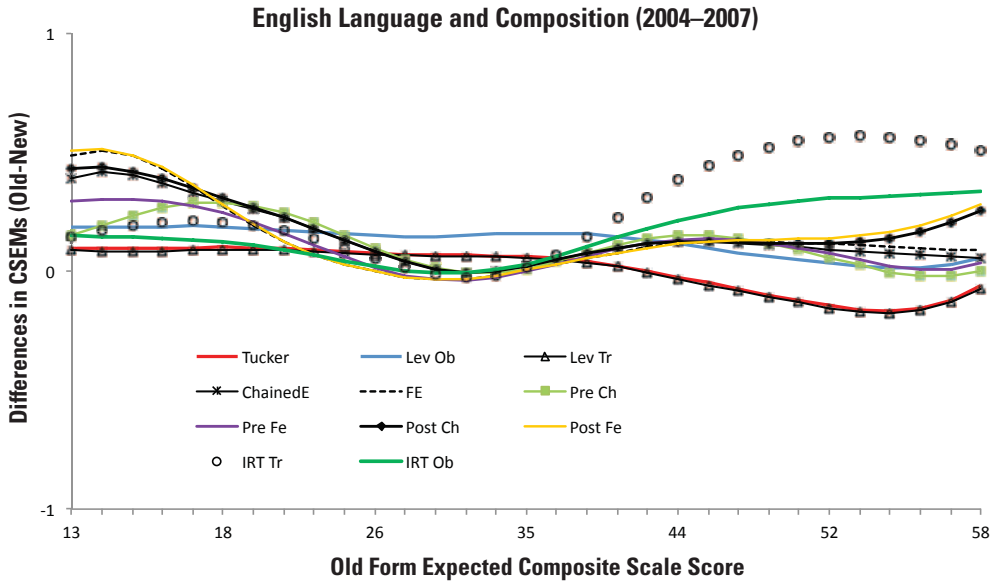


Figure A18.

Differences in CSEMs of composite scale scores for French Language and Culture 2005–2007 (IRT framework, intact forms).

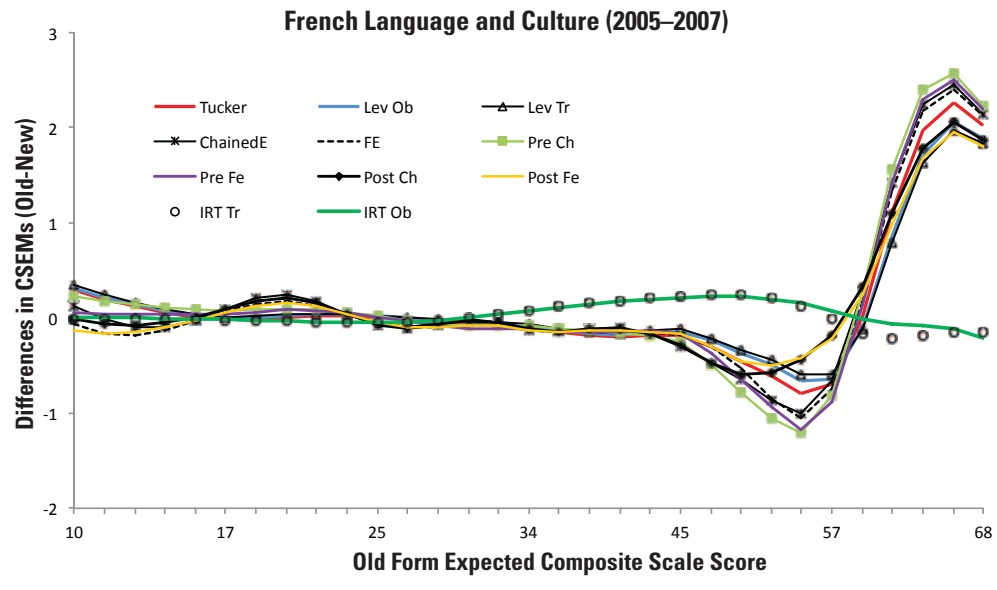


Figure A19.

Differences in expected raw scores for Biology 2004–2006 (BB framework, full-length MC pseudo-tests).

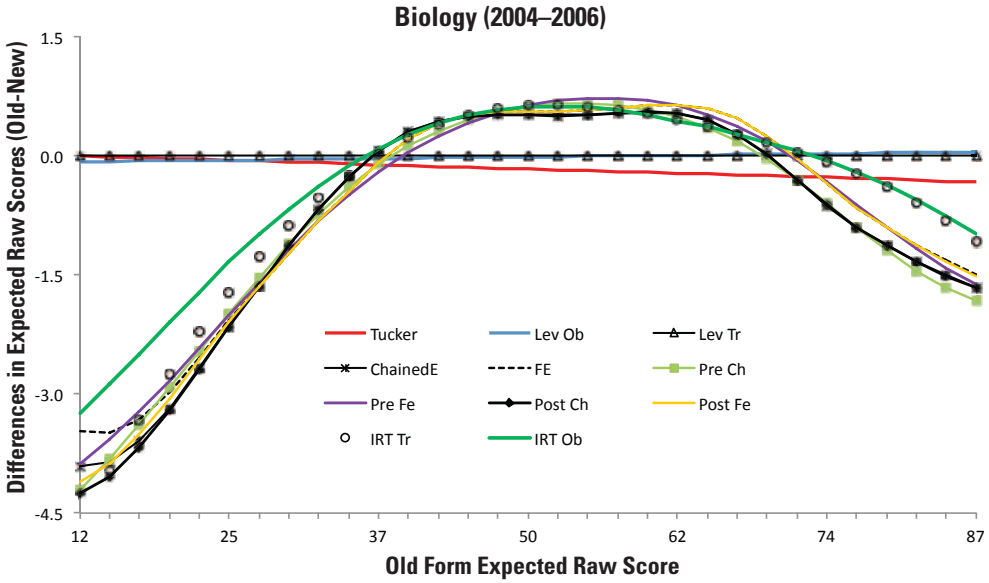


Figure A20.

Differences in expected raw scores for Biology 2004–2006 (IRT framework, full-length MC pseudo-tests).

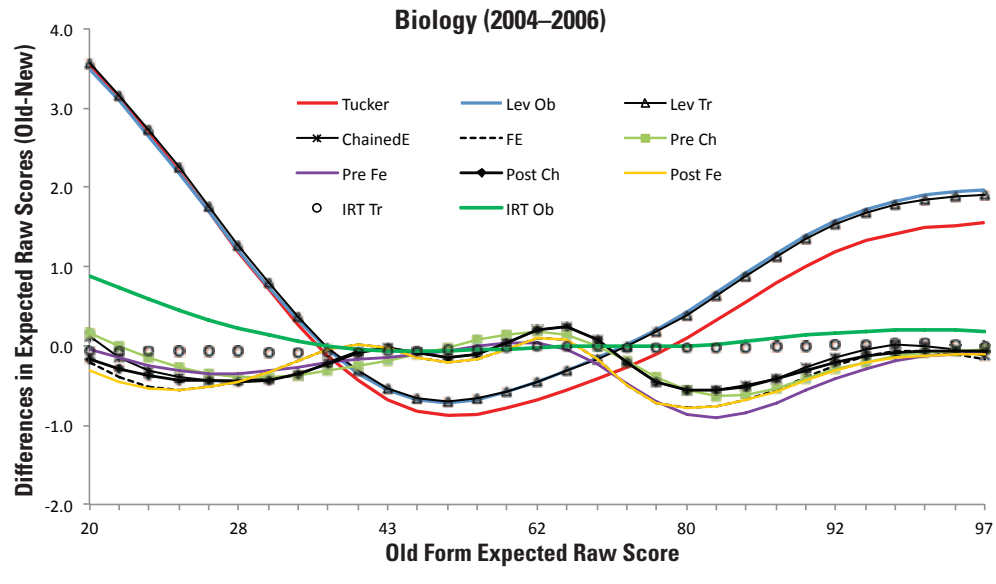


Figure A21.

Differences in expected raw scores for Biology 2005–2007 (BB framework, full-length MC pseudo-tests).

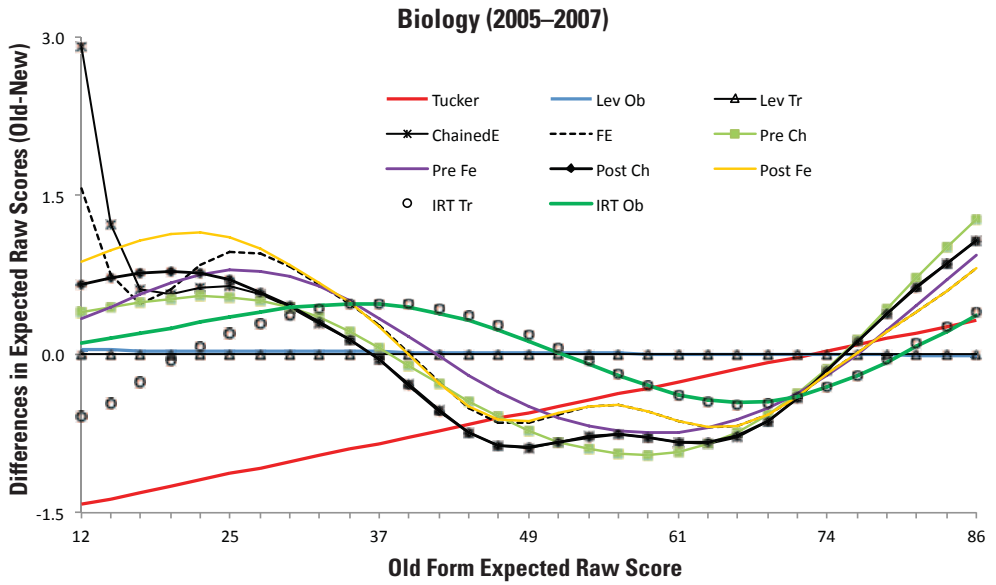


Figure A22.

Differences in expected raw scores for Biology 2005–2007 (IRT framework, full-length MC pseudo-tests).

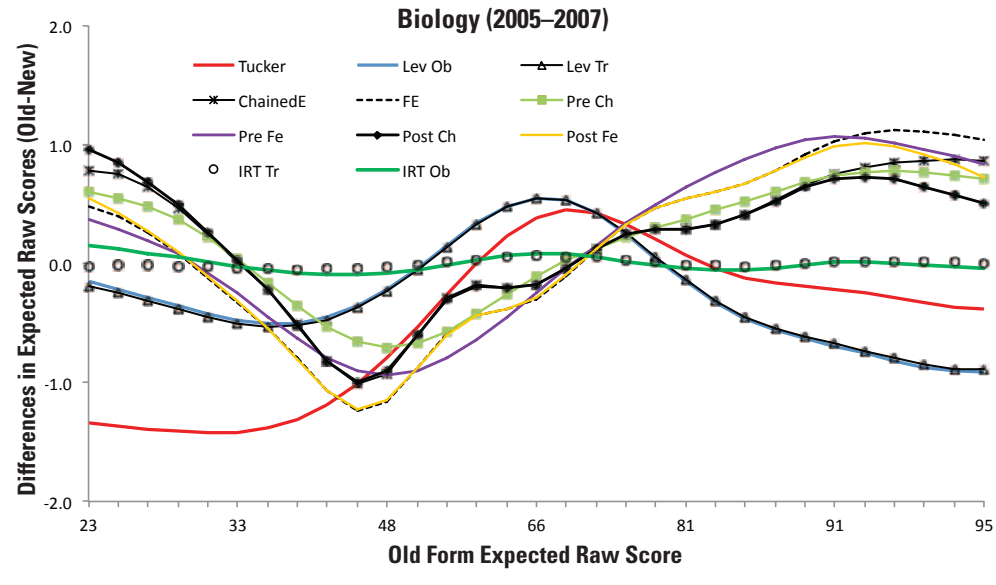


Figure A23.

Differences in expected scale scores for Biology 2004–2006 (BB framework, full-length MC pseudo-tests).

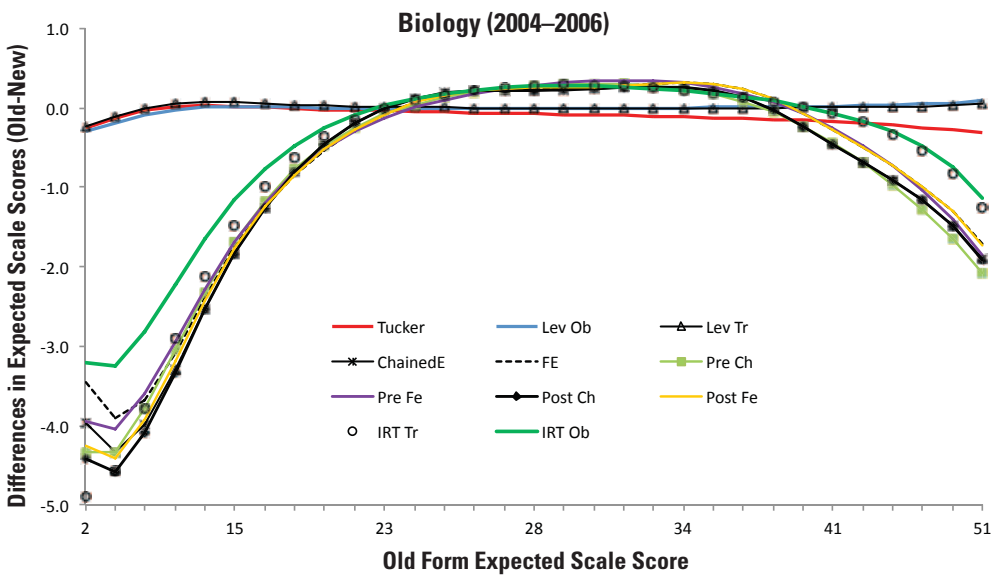


Figure A24.

Differences in expected scale scores for Biology 2004–2006 (IRT framework, full-length MC pseudo-tests).

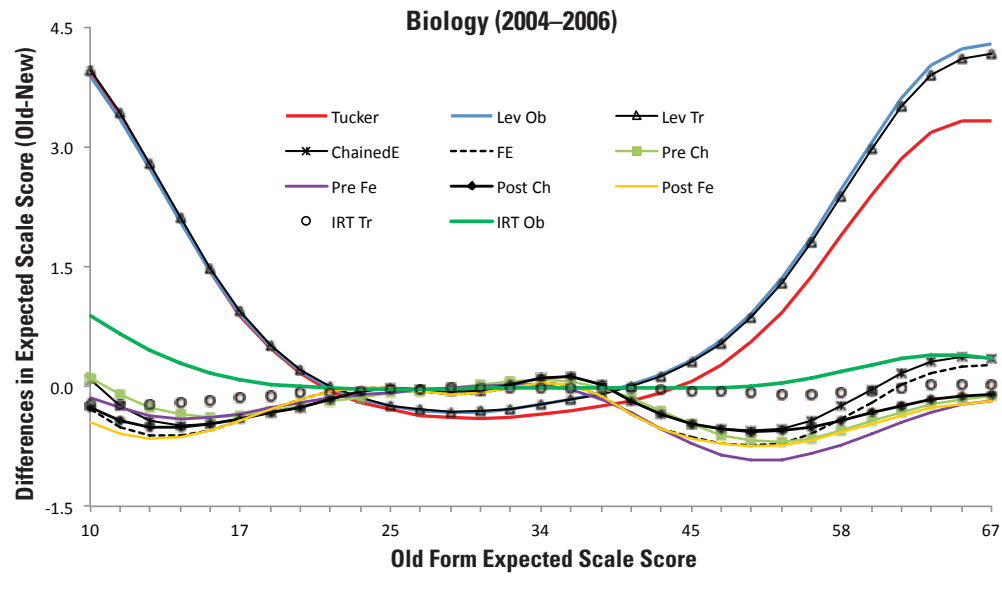


Figure A25.

Differences in expected scale scores for Biology 2005–2007 (BB framework, full-length MC pseudo-tests).

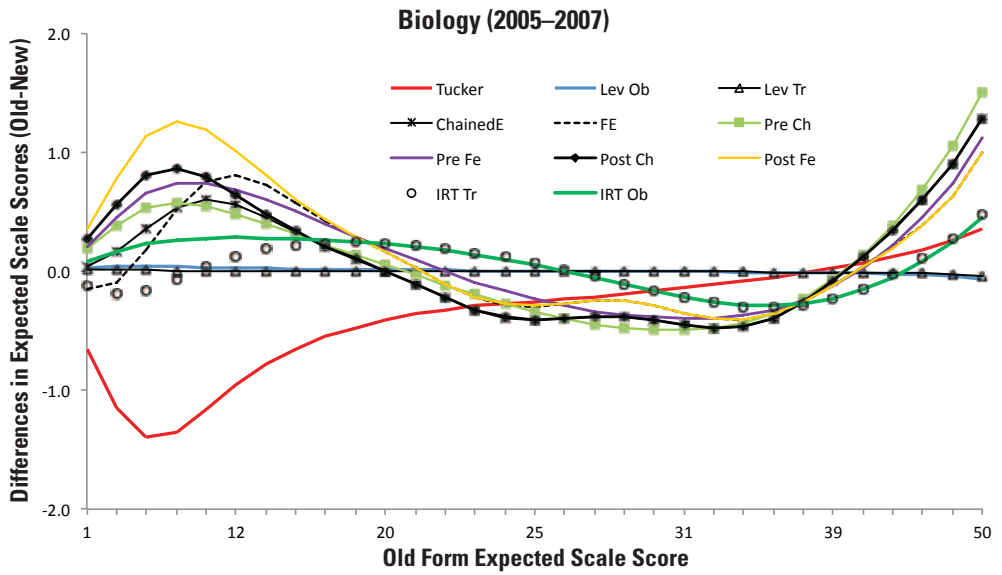


Figure A26.

Differences in expected scale scores for Biology 2005–2007 (IRT framework, full-length MC pseudo-tests).

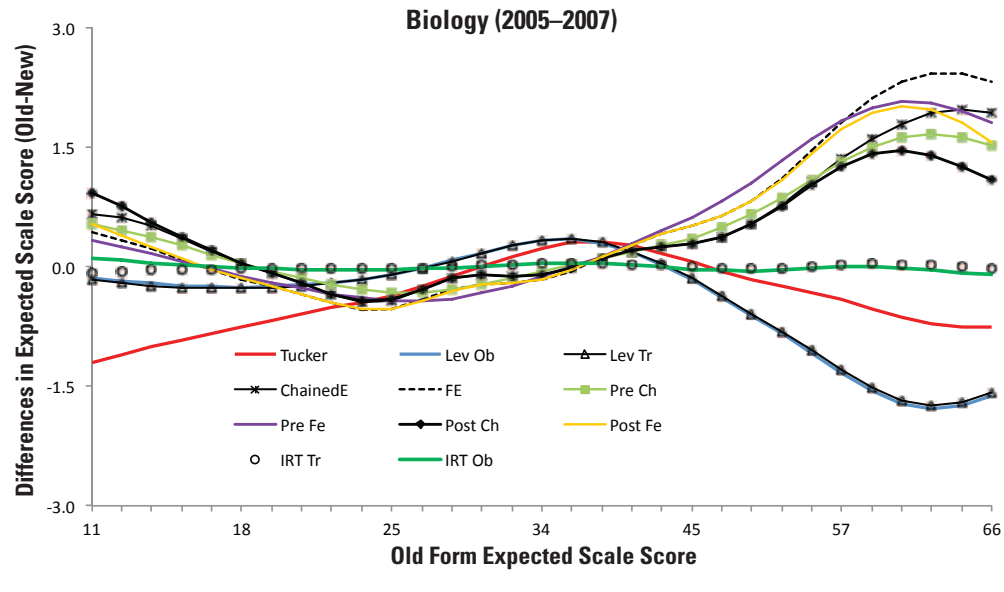


Figure A27.

Differences in CSEMs of raw scores for Biology 2004–2006 (BB framework, full-length MC pseudo-tests).

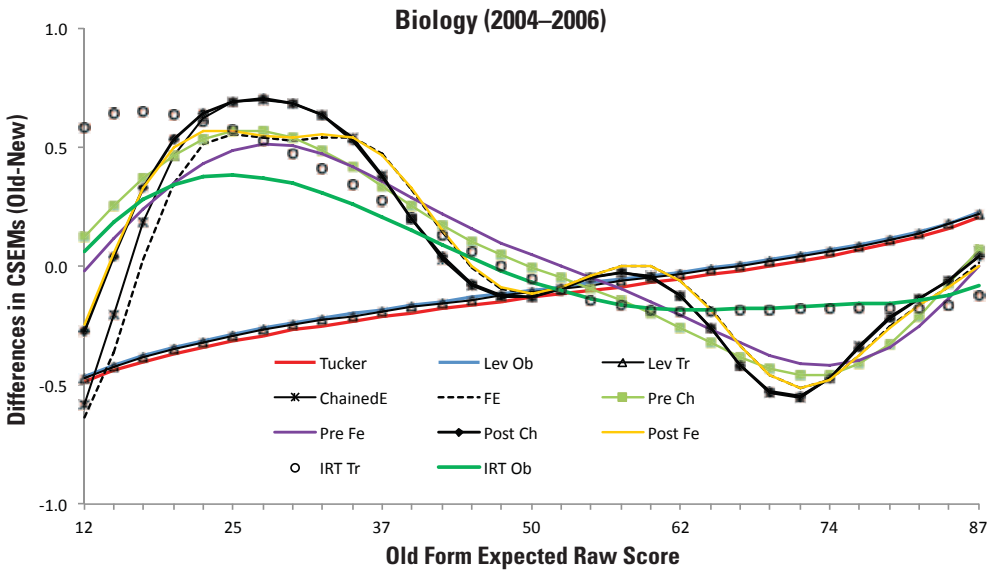


Figure A28.

Differences in CSEMs of raw scores for Biology 2004–2006 (IRT framework, full-length MC pseudo-tests).

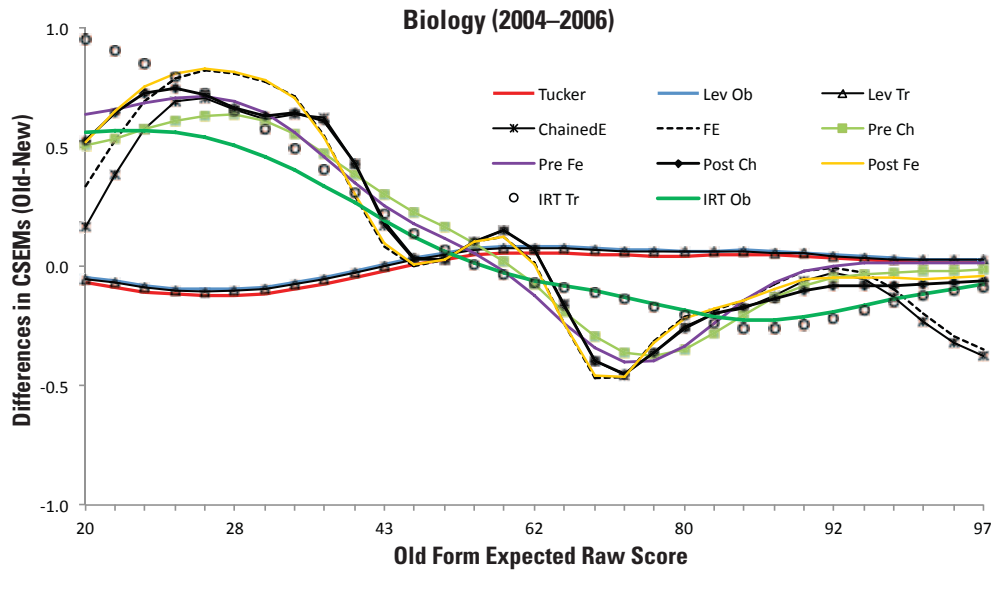


Figure A29.

Differences in CSEMs of raw scores for Biology 2005–2007 (BB framework, full-length MC pseudo-tests).

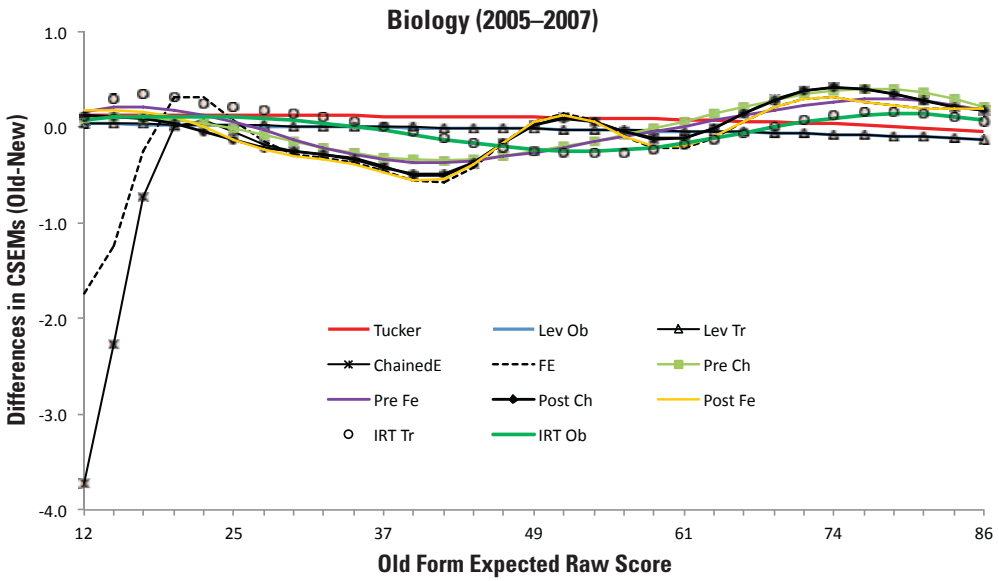


Figure A30.

Differences in CSEMs of raw scores for Biology 2005–2007 (IRT framework, full-length MC pseudo-tests).

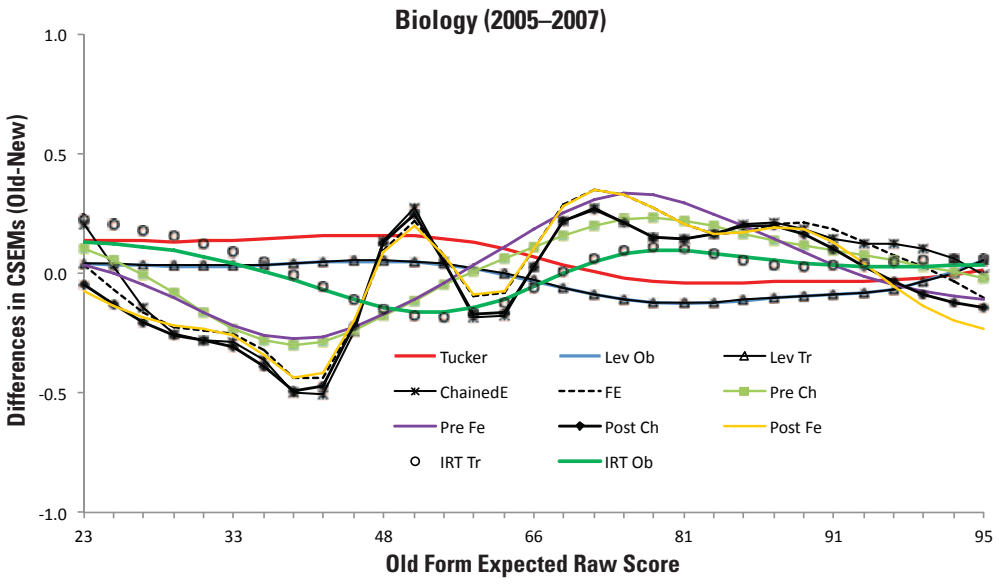


Figure A31.

Differences in CSEMs of scale scores for Biology 2004–2006 (BB framework, full-length MC pseudo-tests).

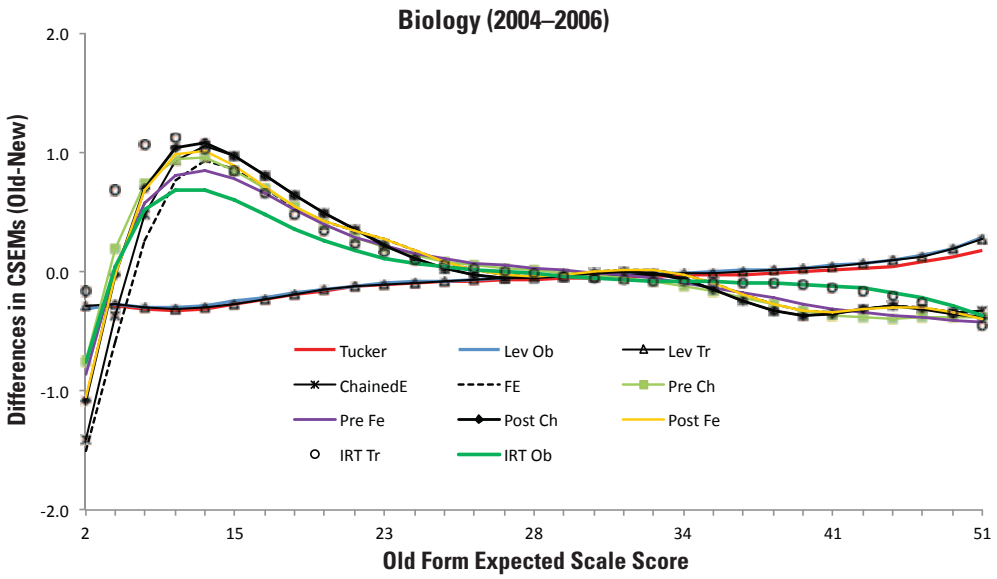


Figure A32.

Differences in CSEMs of scale scores for Biology 2004–2006 (IRT framework, full-length MC pseudo-tests).

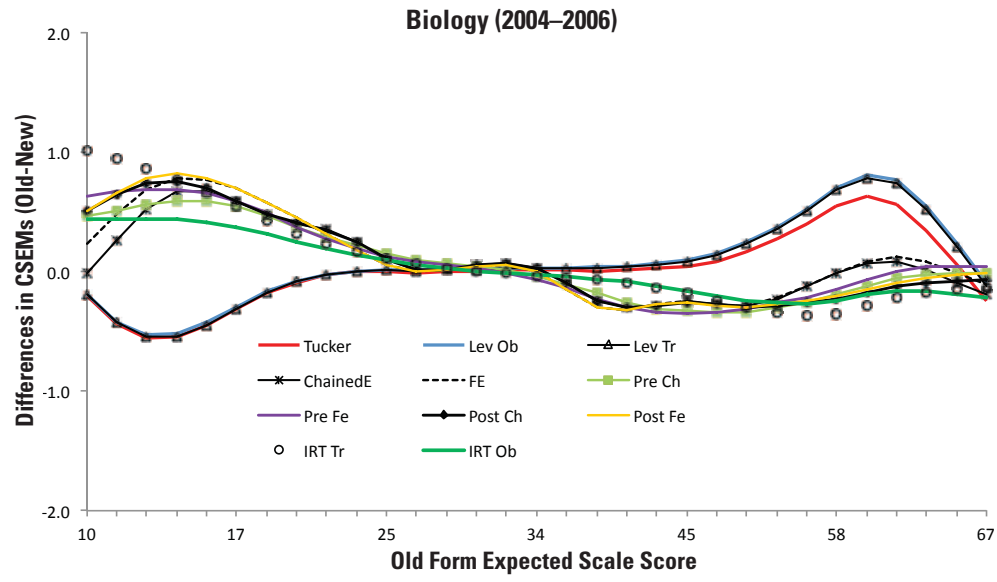


Figure A33.

Differences in CSEMs of scale scores for Biology 2005–2007 (BB framework, full-length MC pseudo-tests).

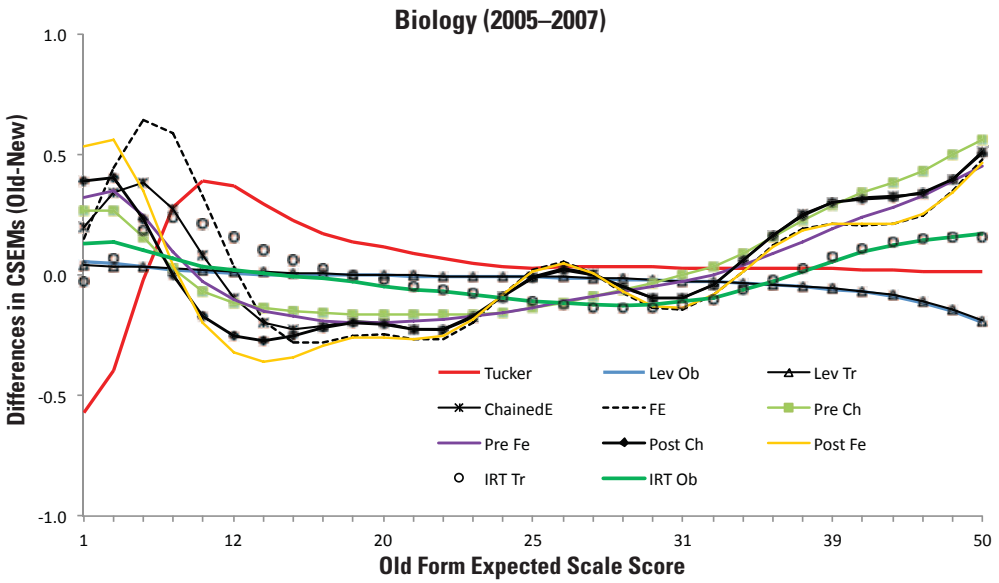


Figure A34.

Differences in CSEMs of scale scores for Biology 2005–2007 (IRT framework, full-length MC pseudo-tests).

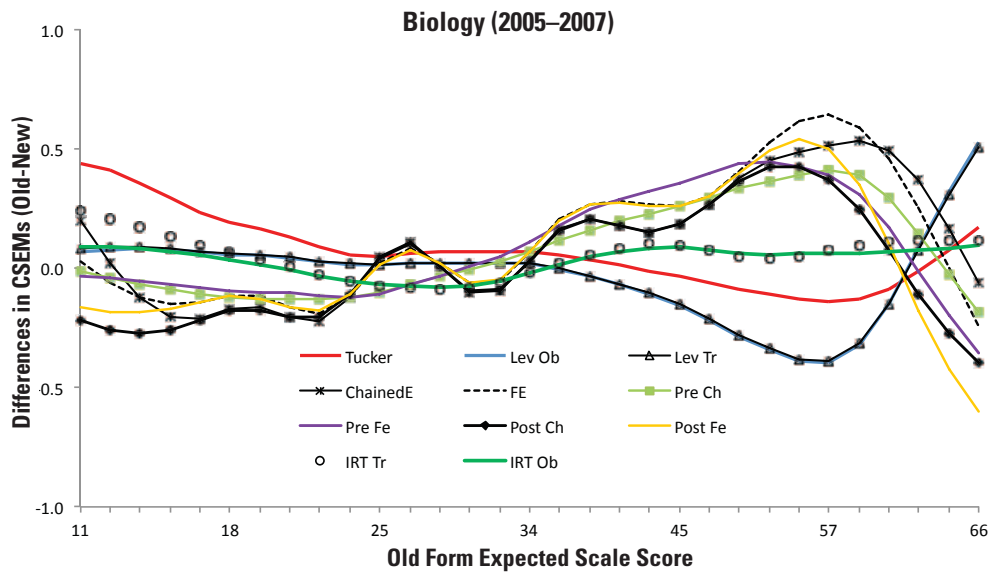


Figure A35.

Differences in expected raw scores for the high-reliability pseudo-test (BB framework).

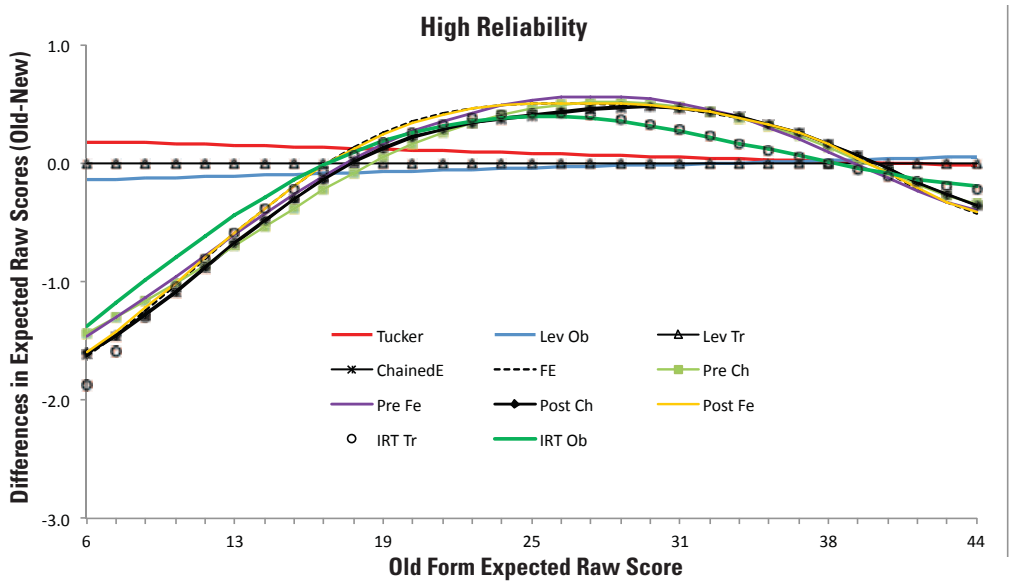


Figure A36.

Differences in expected raw scores for the medium-reliability pseudo-test (BB framework).

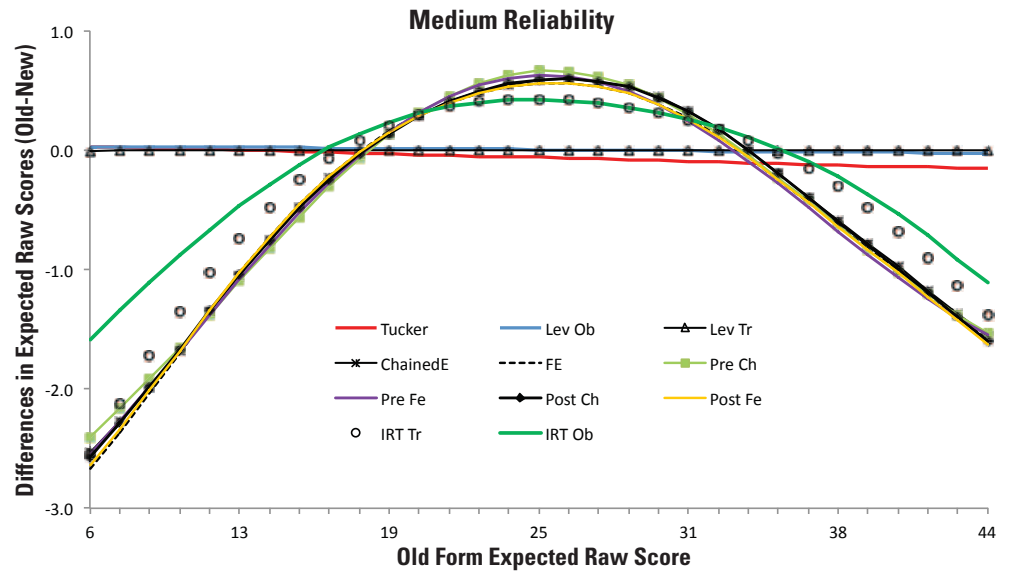


Figure A37.

Differences in expected raw scores for the low-reliability pseudo-test (BB framework).

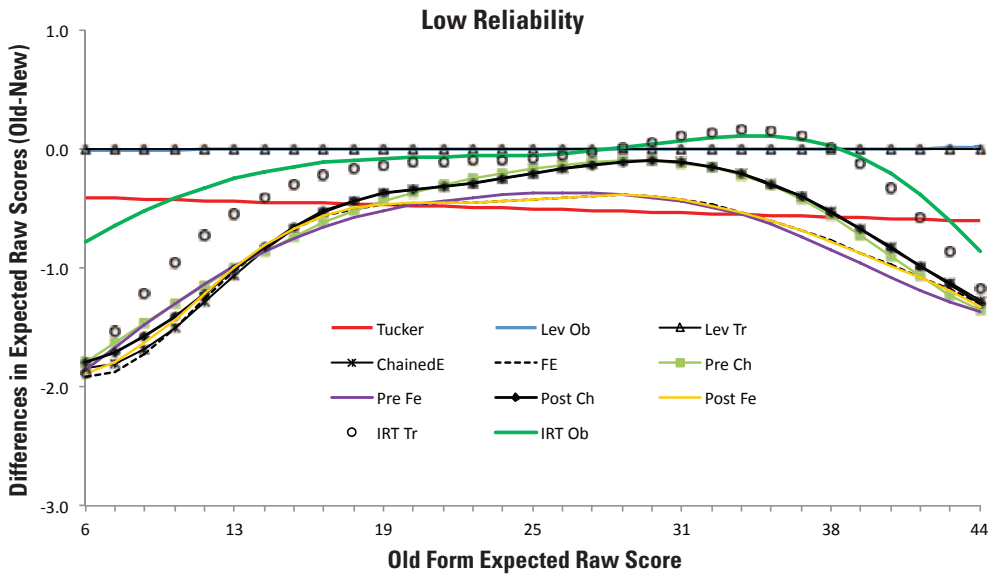


Figure A38.

Differences in expected raw scores for the high-reliability pseudo-test (IRT framework).

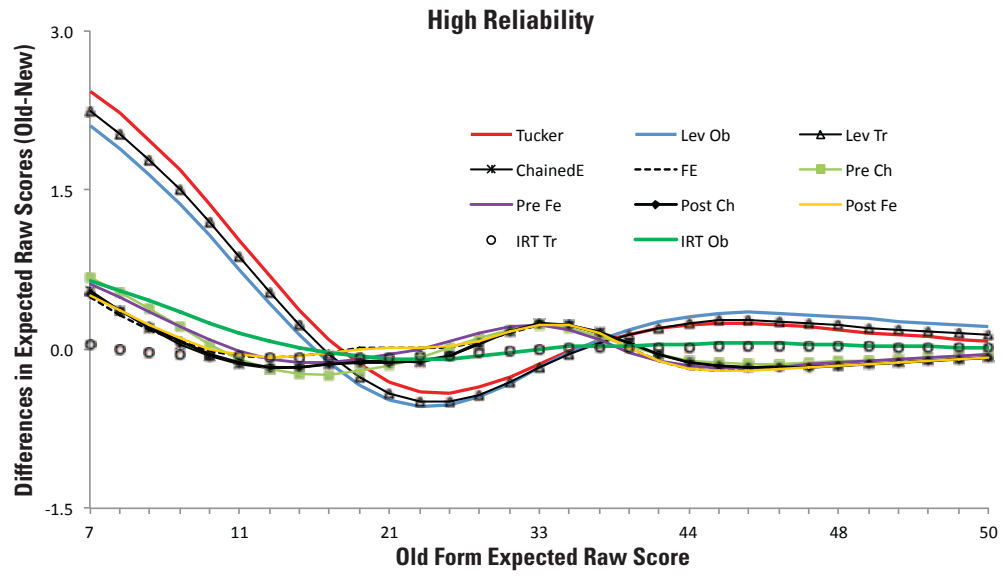


Figure A39.

Differences in expected raw scores for medium-reliability pseudo-test (IRT framework).

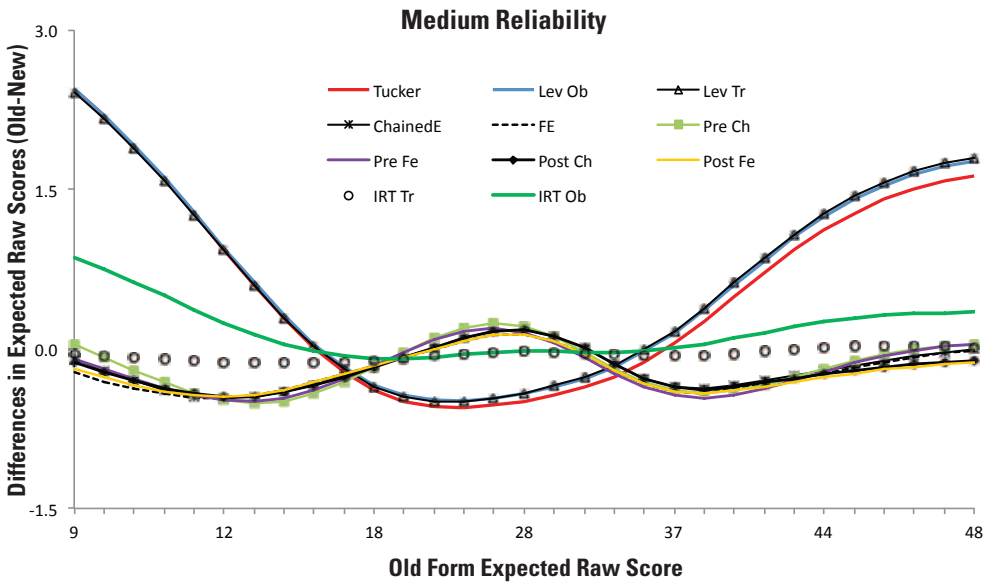


Figure A40.

Differences in expected raw scores for low-reliability pseudo-test (IRT framework).

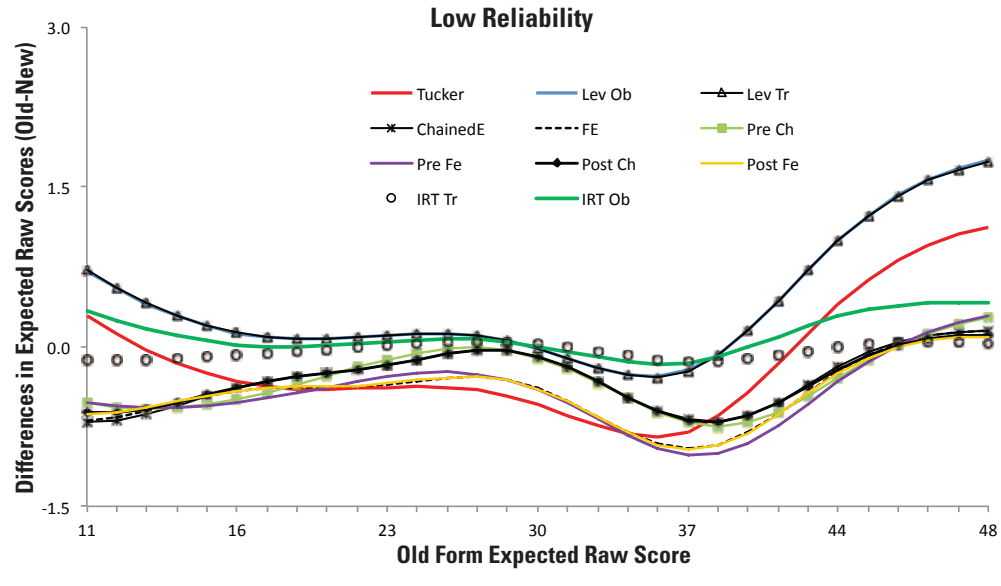


Figure A41.

Differences in expected scale scores for the high-reliability pseudo-test (BB framework).

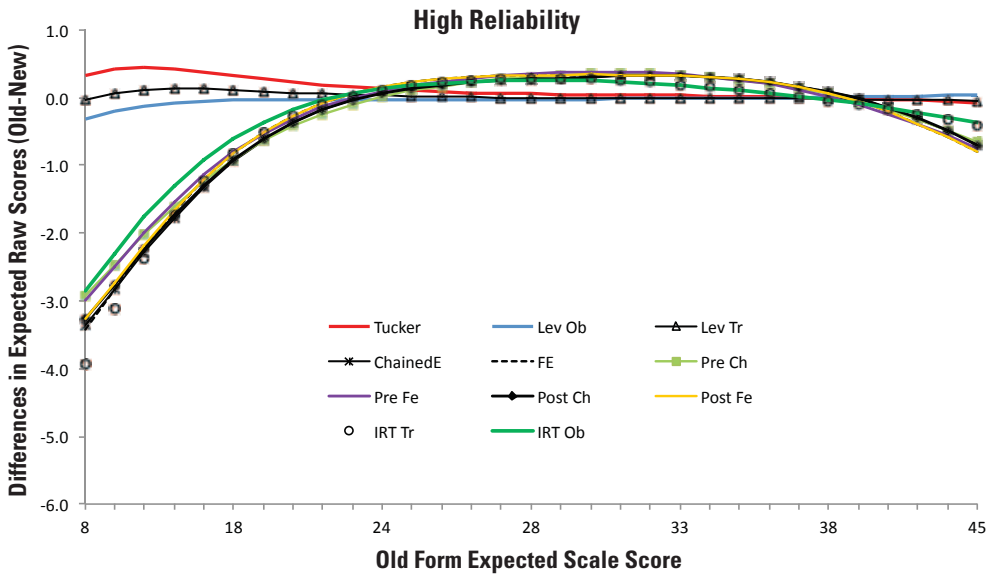


Figure A42.

Differences in expected scale scores for the medium-reliability pseudo-test (BB framework).

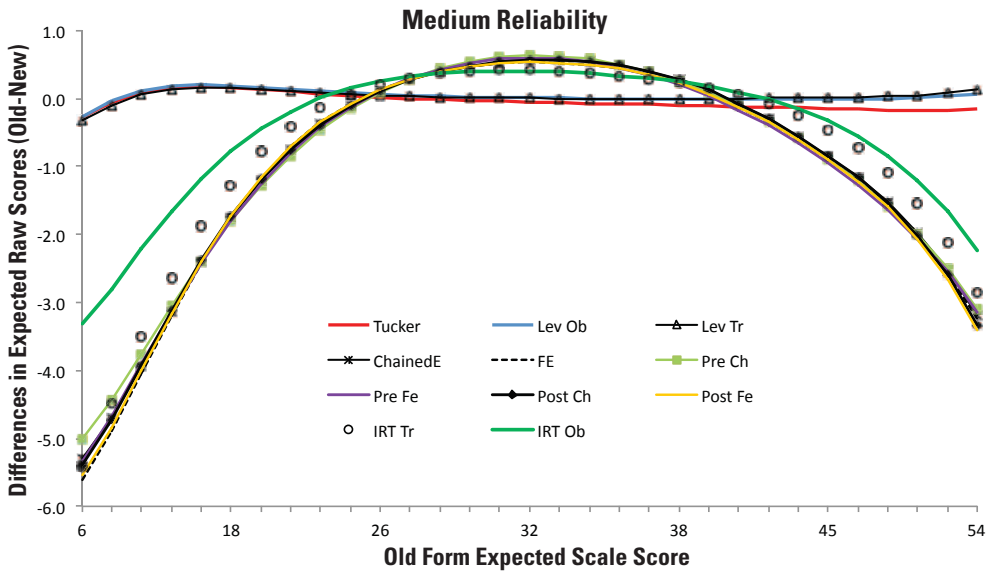


Figure A43.

Differences in expected scale scores for the low-reliability pseudo-test (BB framework).

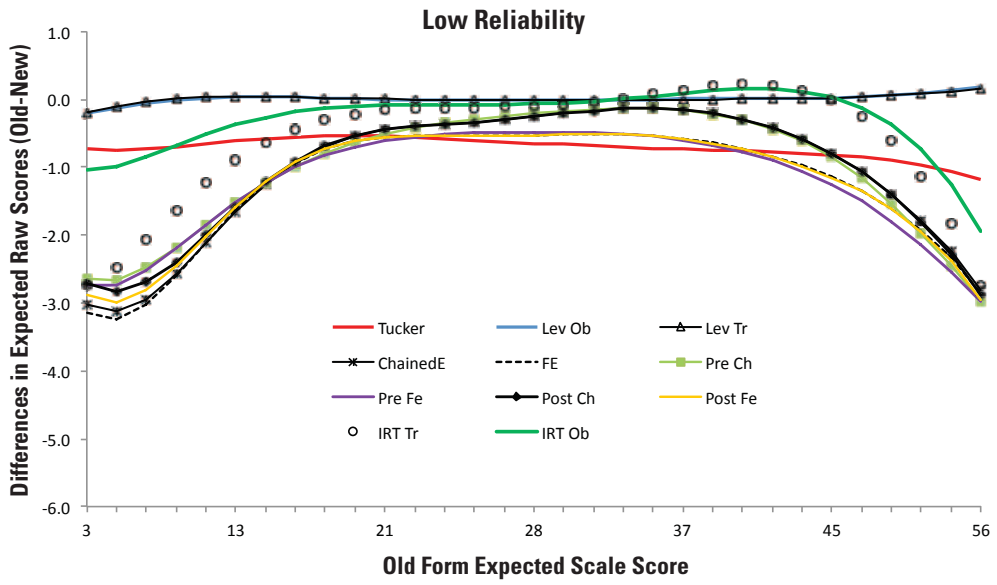


Figure A44.

Differences in expected scale scores for the high-reliability pseudo-test (IRT framework).

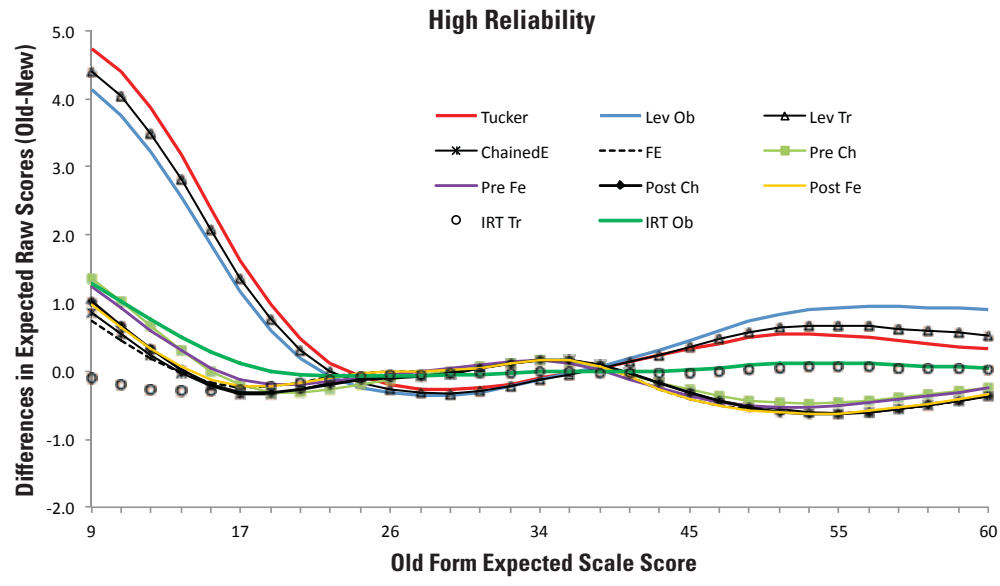


Figure A45.

Differences in expected scale scores for the medium-reliability pseudo-test (IRT framework).

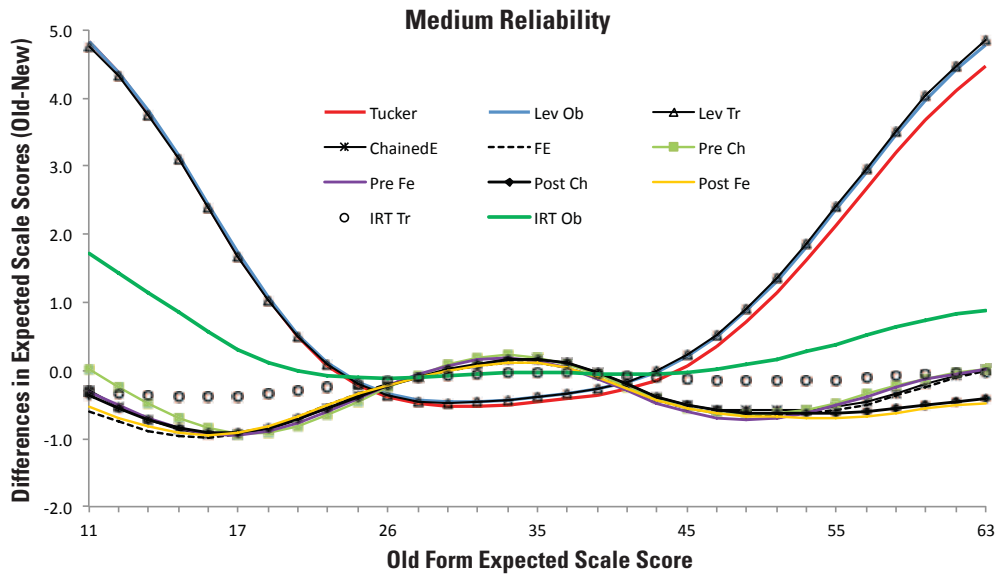


Figure A46.

Differences in expected scale scores for the low-reliability pseudo-test (IRT framework).

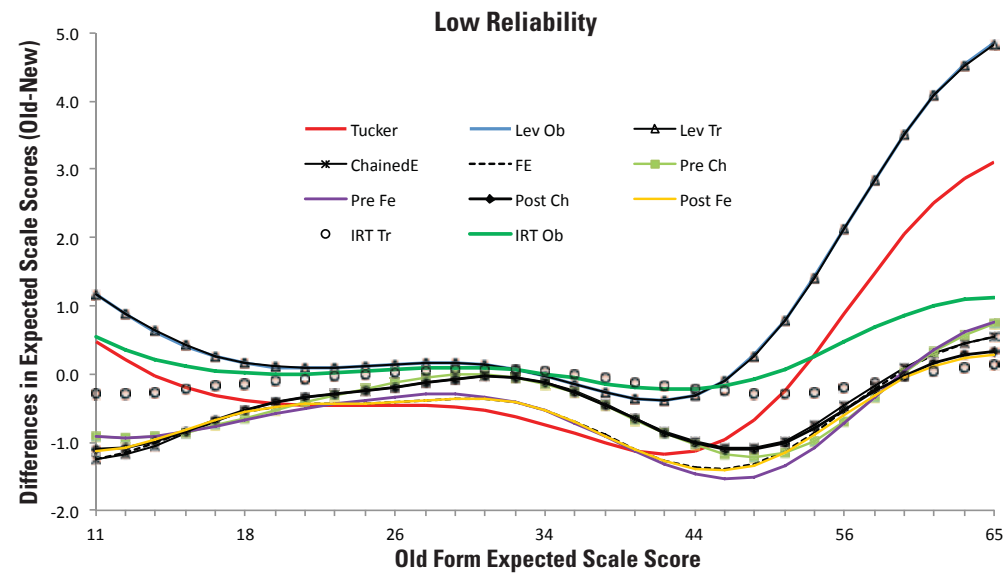


Figure A47.

Differences in CSEMs of raw scores for the high-reliability pseudo-test (BB framework).

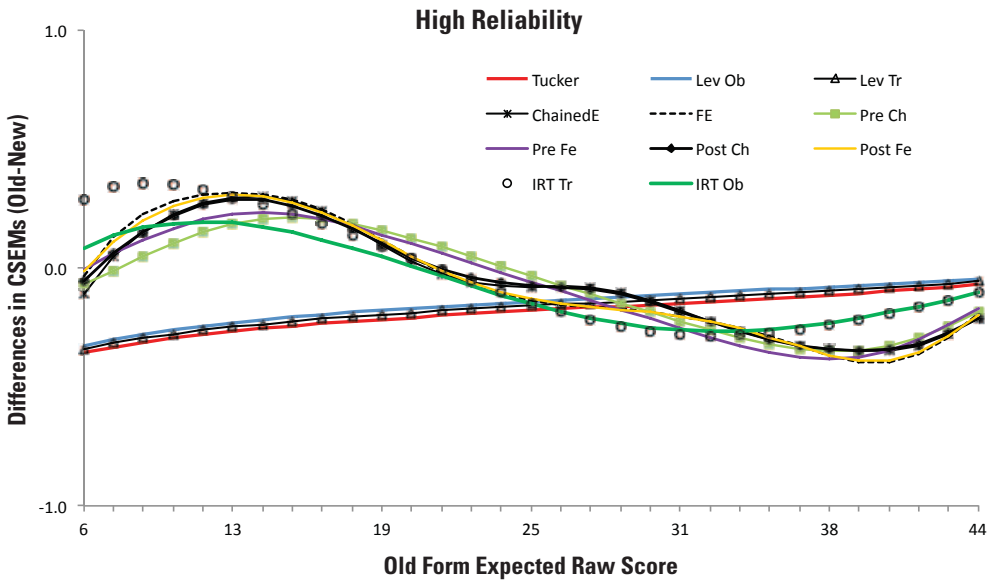


Figure A48.

Differences in CSEMs of raw scores for the medium-reliability pseudo-test (BB framework).

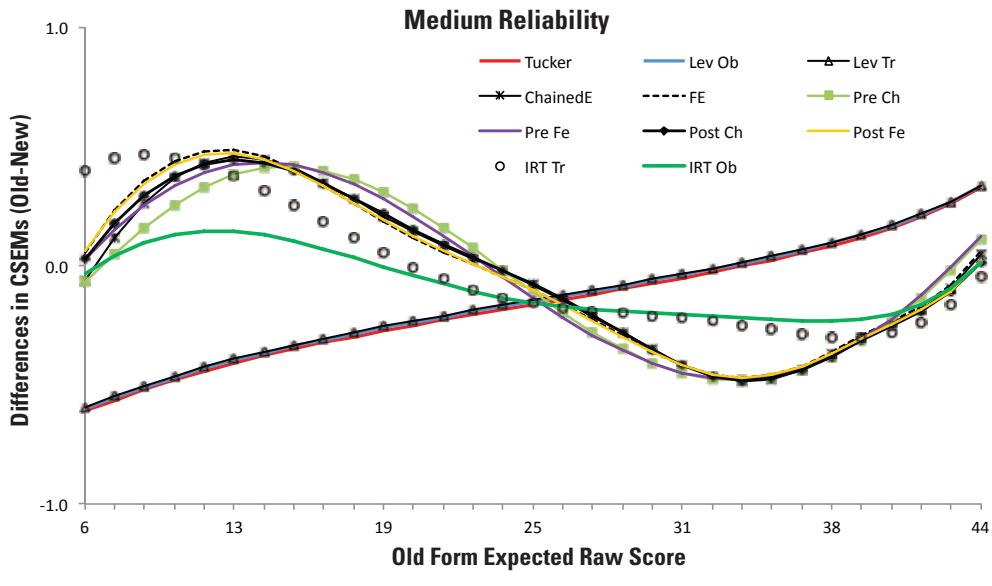


Figure A49.

Differences in CSEMs of raw scores for the low-reliability pseudo-test (BB framework).

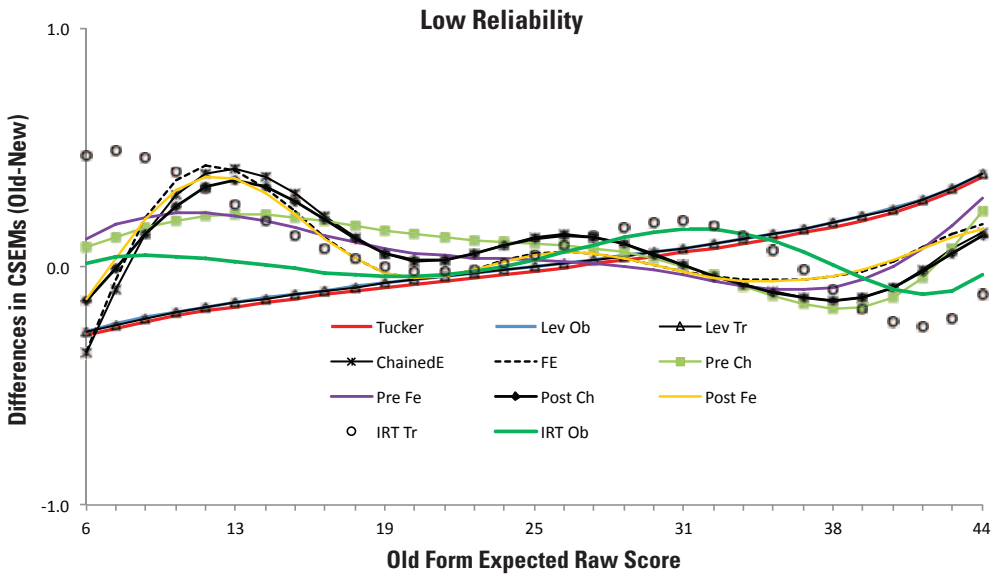


Figure A50.

Differences in CSEMs of raw scores for the high-reliability pseudo-test (IRT framework).

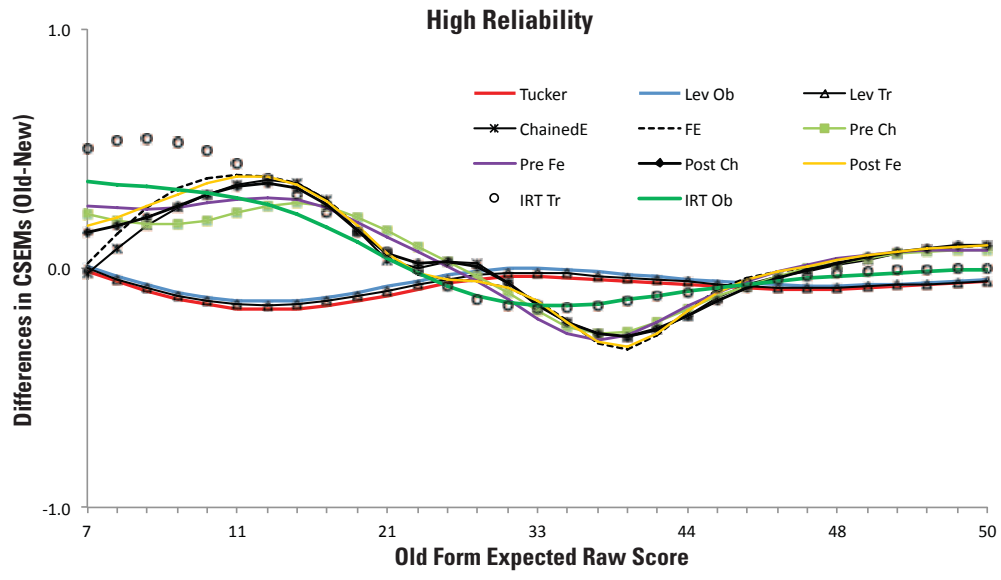


Figure A51.

Differences in CSEMs of raw scores for the medium-reliability pseudo-test (IRT framework).

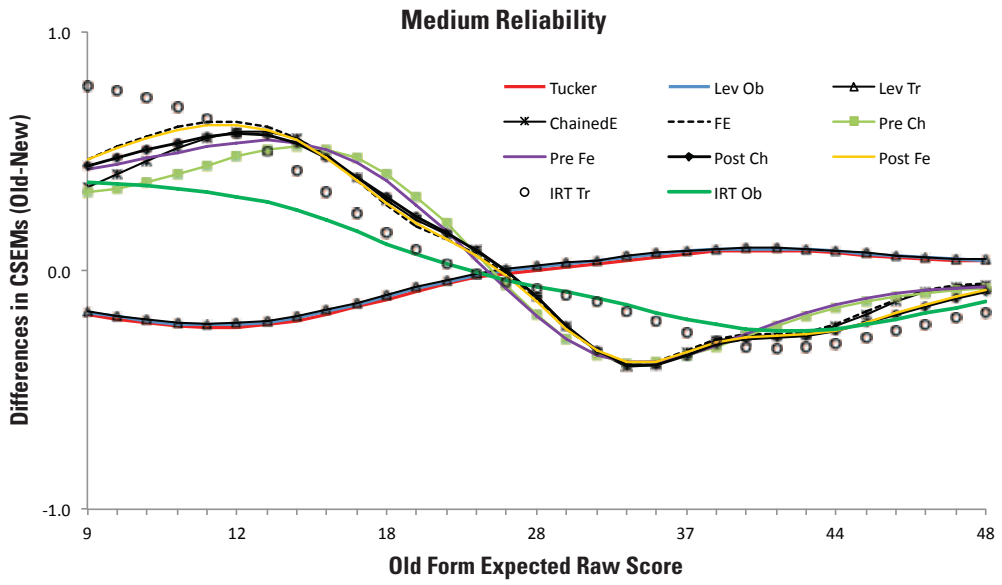


Figure A52.

Differences in CSEMs of raw scores for the low-reliability pseudo-test (IRT framework).

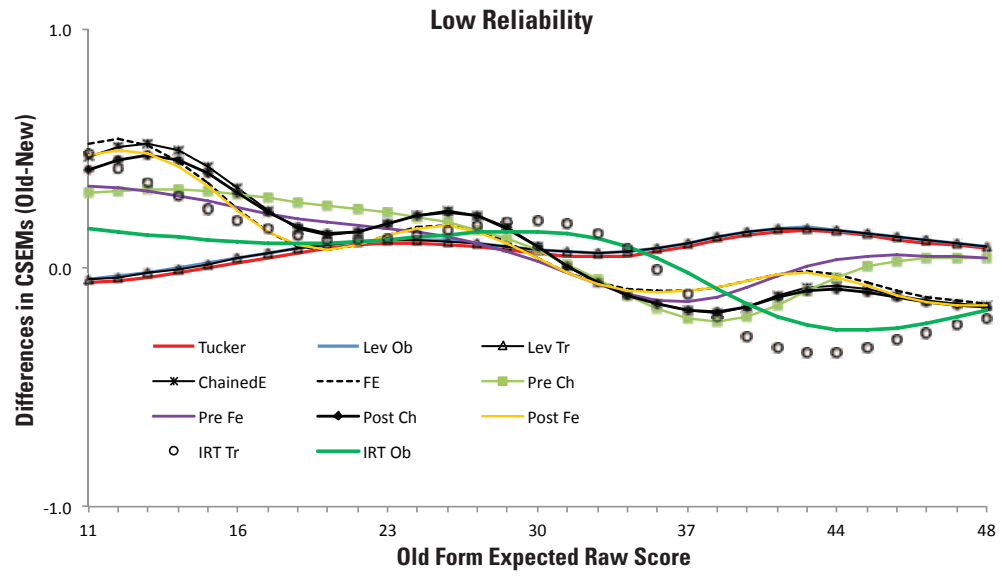


Figure A53.

Differences in CSEMs of raw scores for the high-reliability pseudo-test (BB framework).

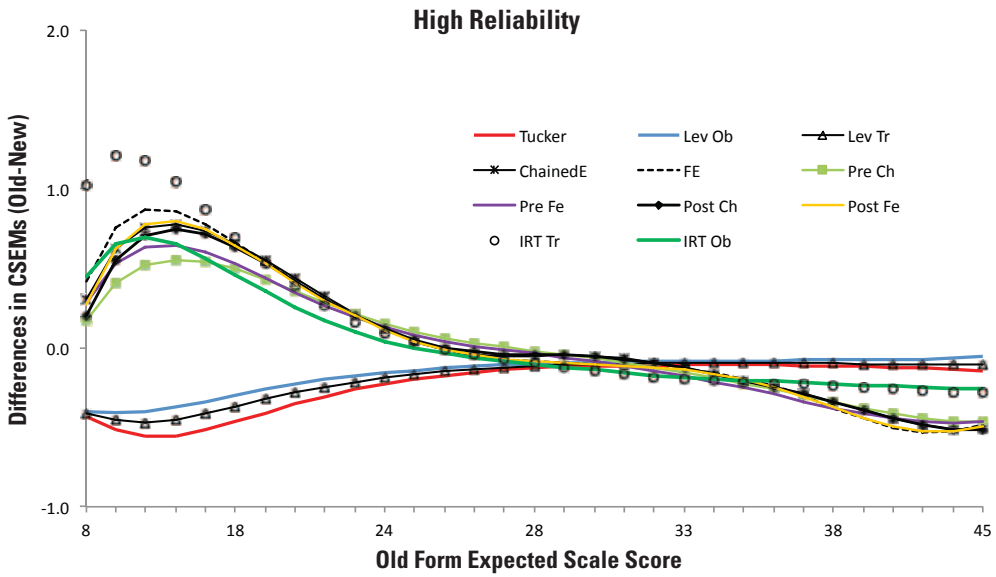


Figure A54.

Differences in CSEMs of scale scores for the medium-reliability pseudo-test (BB framework).

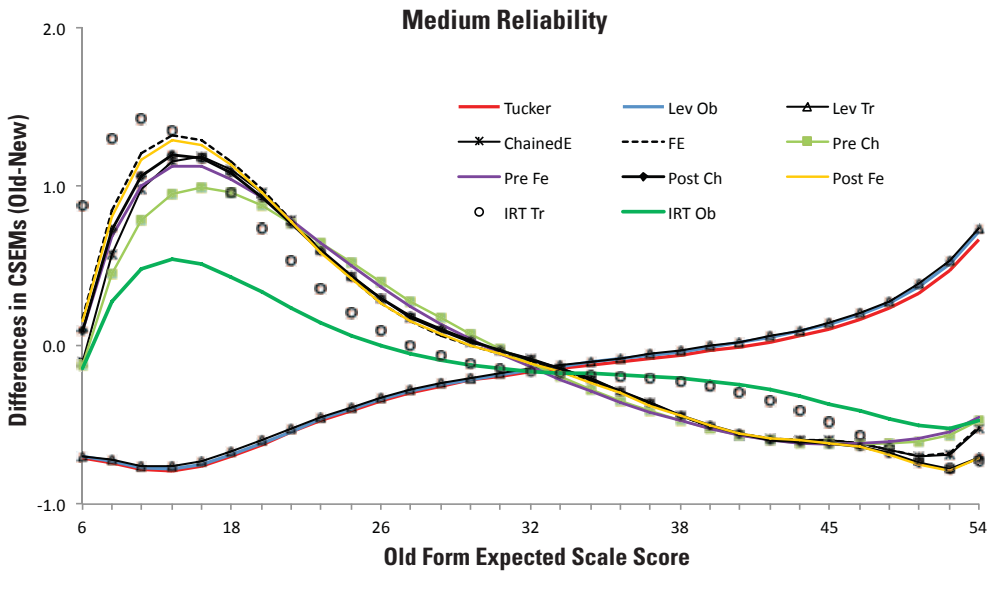


Figure A55.

Differences in CSEMs of scale scores for the low-reliability pseudo-test (BB framework).

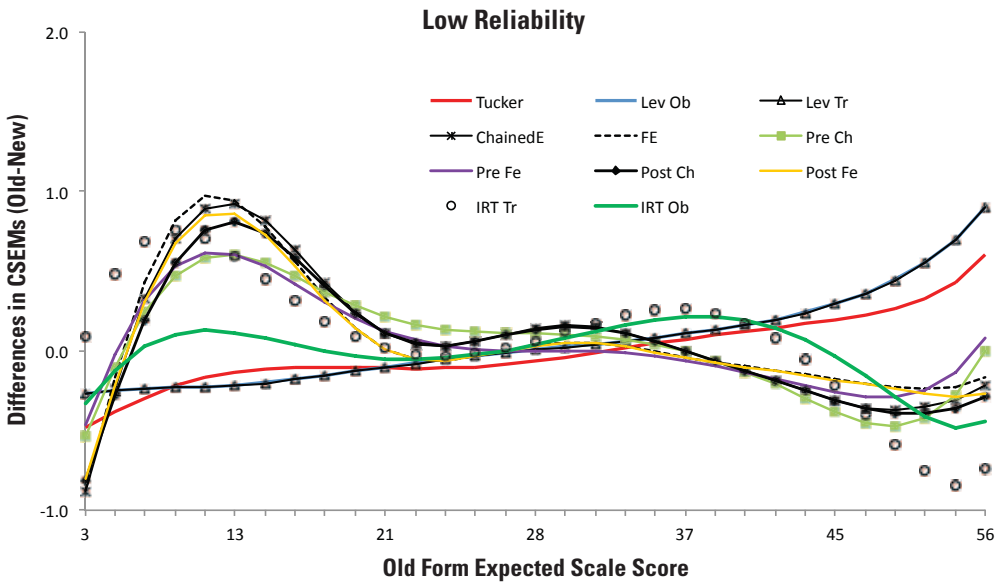


Figure A56.

Differences in CSEMs of scale scores for the high-reliability pseudo-test (IRT framework).

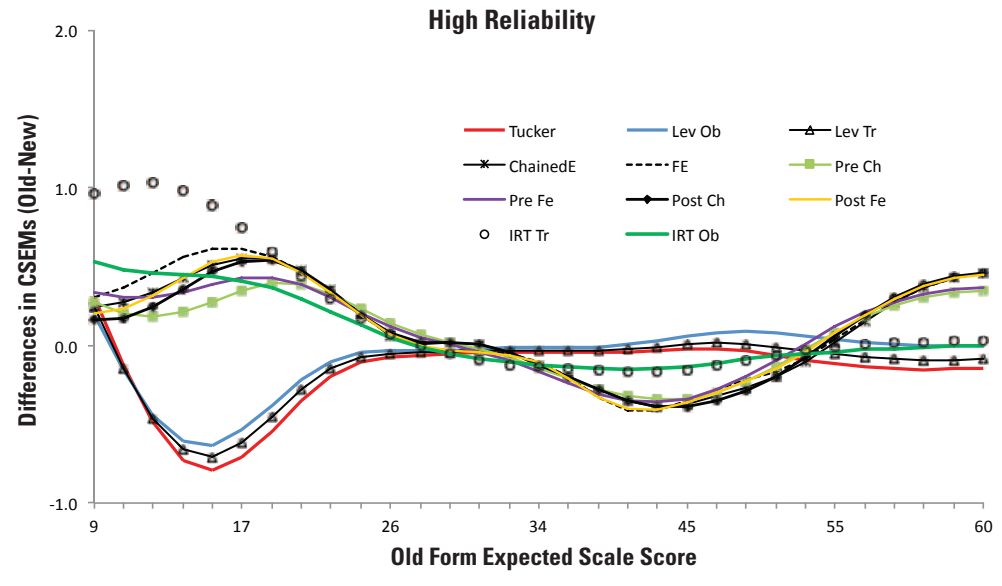


Figure A57.

Differences in CSEMs of scale scores for the medium-reliability pseudo-test (IRT framework).

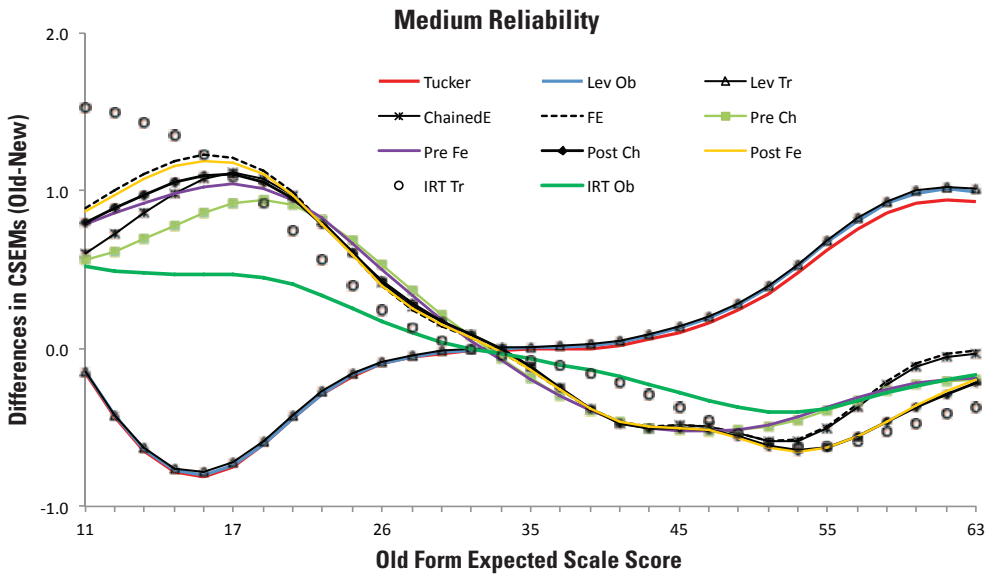
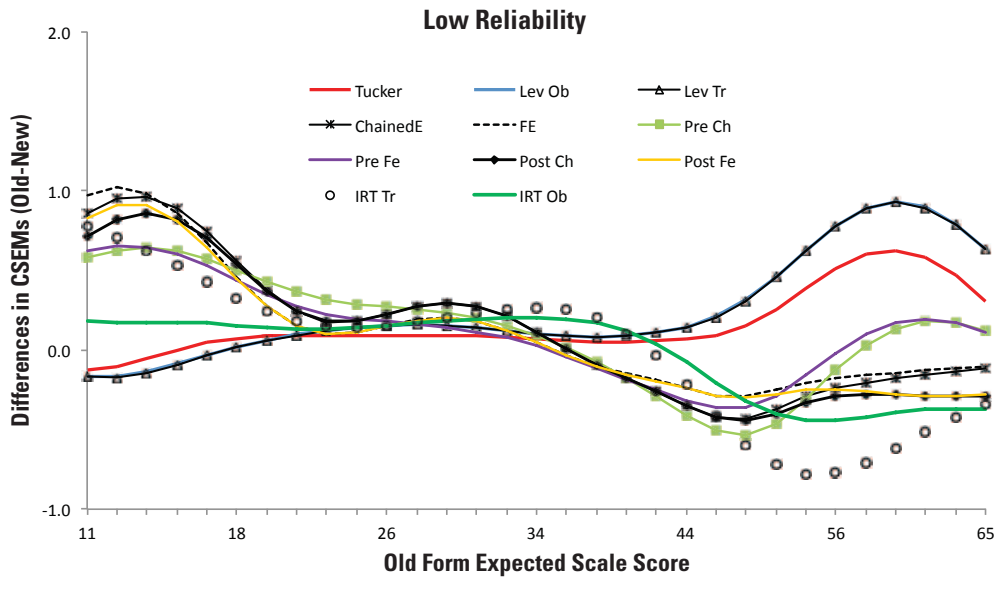


Figure A58.

Differences in CSEMs of scale scores for the low-reliability pseudo-test (IRT framework).



The Research & Development 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®
- Developing standards and conducting college and career readiness alignment studies
- 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: collegeboard.org/research

CollegeBoard

CollegeBoard

12b-5133