Center for Advanced Studies in
Measurement and Assessment

CASMA Monograph

Number 2.3

Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 3)

Michael J. Kolen
Won-Chan Lee
(Editors)

December, 2014

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Preface for Volume 3

This monograph, *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 3*), continues the work presented in Volume 1 (Kolen & Lee, 2011) and Volume 2 (Kolen & Lee, 2012). As stated in the Preface of the first volume,

> Beginning in 2007 and continuing through 2011, with funding from the College Board, we initiated a research program to investigate psychometric methodology for mixed-format tests through the Center for Advanced Studies in Measurement and Assessment (CASMA) at the University of Iowa. This research uses data sets from the Advanced Placement (AP) Examinations that were made available by the College Board. The AP examinations are mixed-format examinations that contain multiple-choice (MC) and a substantial proportion of free response (FR) items. Scores on these examinations are used to award college credit to high school students who take AP courses and earn sufficiently high scores. There are more than 30 AP examinations.
>
> We had two major goals in pursuing this research. First, we wanted to contribute to the empirical research literature on psychometric methods for mixed-format tests, with a focus on equating. Second, we wanted to provide graduate students with experience conducting empirical research on important and timely psychometric issues using data from an established testing program.

Refer to the Preface for Volume 1 for more background on this research. The work on Volume 2, completed in 2012, extended the work from Volume 1. The work in Volume 3, was also supported with funding from the College Board, and provides a further extension.

Volume 3 contains 8 chapters. Chapter 1 provides an overview. In addition, it highlights some of the methodological issues encountered and some of the major findings.

Chapter 2, which is an extension of Chapter 2 in Volume 2, is a simulation study that investigates the feasibility of equating mixed-format test forms using a common-item set that consists solely of MC items. Chapter 3 is also a simulation study that compares the equating of mixed-format tests using common-item sets that contain solely of MC items to common-item sets that contain both MC and FR items.

Chapter 4 compares, for mixed-format tests, item response theory (IRT) item parameter estimates from four IRT calibration software programs using a real data set. Chapter 5 examines the influence of IRT calibration programs on IRT equating results for mixed-format tests using some of the same data sets used in Chapter 4.

Chapter 6 compares various factor analysis-based dimensionality assessment procedures using data from four different mixed-format examinations. Chapter 7 compares results from equating mixed-

format tests using a bi-factor multidimensional IRT and unidimensional IRT methods. Chapter 8 examines multidimensional IRT equating methods and compares results from multidimensional IRT equating methods to unidimensional IRT and traditional equating methods.

We thank Guemin Lee (Professor at Yonsei University), Jaime Peterson (now at Pearson), Wei Wang (now at Educational Testing Service), and current graduate students Seohong Pak, Shichao Wang, and Mengyao Zhang for their work. We also thank Shichao Wang and Yujin Kang for their effort in producing this volume.

We thank Robert L. Brennan who provided us with guidance where needed, and, who, as the Co-Director of CASMA, provided us with a home to conduct this work. Thanks to Jennifer Jones and Anne Wilson for their administrative work. Also, we would like to recognize the continuing support provided by Dean's office of the College of Education. We are especially appreciative of the substantial support provided by the College Board as well as College Board staff Kevin Sweeney, Amy Hendrickson, Gerald Melican, Rosemary Reshetar, and Pamela Kaliski.

Michael J. Kolen
Won-Chan Lee
December, 2014
Iowa City, Iowa

References

Kolen, M. J., & Lee, W. (Eds.). (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1). (CASMA Monograph Number 2.1)*. Iowa City, IA: CASMA, The University of Iowa.

Kolen, M. J., & Lee, W. (Eds.). (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2). (CASMA Monograph Number 2.2)*. Iowa City, IA: CASMA, The University of Iowa.

Contents

Chapter 1: Introduction and Overview for Volume 3

Michael J. Kolen and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

This chapter provides an overview of this volume. It describes relationships between the chapters in Volume 3 to those in Volume 1 and Volume 2. In addition, this chapter highlights some of the major findings. This chapter begins with a description of the other chapters. The findings from Volume 3 are related to findings from the earlier volumes. The chapter concludes with a brief discussion.

Introduction and Overview for Volume 3

The research described in Volume 3 is closely related to the research conducted in Volume 1 (Kolen & Lee, 2011) and Volume 2 (Kolen & Lee, 2012). This chapter provides an overview of Volume 3. It describes relationships between the chapters in Volume 3 to those in Volume 1 and Volume 2. In addition, this chapter highlights some of the major findings.

Although the research in this monograph was conducted using data from the Advanced Placement (AP) Examinations, the data were manipulated in such a way that the research does not pertain directly to operational AP examinations. Instead, it is intended to address general research questions that would be of interest in many testing programs.

The chapter begins with a description of the research questions, designs, and findings. The findings from Volume 3 are related to findings from the earlier volumes. The chapter concludes with a brief discussion.

Research Summary

Chapters 2 and 3 investigate, through simulation, the effects of the composition of common-item sets on mixed-format test equating. These studies are direct extensions of studies conducted in the first two volumes. Chapters 4 through 8 investigate various issues related to item response theory (IRT) equating with mixed-format tests. Chapter 4 compares IRT item parameter estimates from four different software programs. Chapter 5 examines the effect on IRT equating of differences in item parameter estimates from the software programs studied in Chapter 4. Chapter 6 investigates various factor analysis-based methods for assessing dimensionality. Chapters 7 and 8 investigate multidimensional IRT equating methods. Research questions, designs, and findings are summarized in this section.

Chapter 2

Chapter 2 is a direct extension of the Lee, He, Hagge, Wang, and Kolen (2012) simulation study from Volume 2. These studies investigated the conditions under which scores on mixed-format test forms can be adequately equated using a set of common items that consists solely of multiple-choice (MC) items. In Chapter 2 the following variables were manipulated: sample size (3 levels); traditional equating method (12 methods); group proficiency differences, including conditions where the group ability differences were not the same for the free-response (FR) and MC items (9 levels); correlation between MC and FR items (9 levels); length of common item set (2 levels); and score scale (3 types). One of novel findings of this study is that

equating contains substantial error when the correlation between scores on MC and FR items differs for the two forms to be equated. It was also found that large group differences for FR items introduced substantial equating error. Similar to other studies, it was found that there was less equating error as sample size increases, length of common item set increases, and group proficiency difference decreases.

Chapter 3

For the simulation study in Chapter 3, pools of items were assembled using item parameter estimates from various AP examinations. These pools were used to construct simulated test forms with various characteristics. The focus of this study was to compare, on the adequacy of equating, the use of MC-only sets of common items with sets of common items that contain both MC and FR items. The variables manipulated in this study included item-type multidimensionality, group ability difference, and equating method. Overall, it was found that the MC-only common item set produced acceptable amounts of equating error when using chained equipercentile methods and group ability differences are not large.

Chapter 4

In Chapter 4, item parameter estimation for a mixed-format test was compared across four software programs and four combinations of IRT models for the MC and FR items. Analyses were completed on a single data set. Differences in item parameter estimates, test characteristic curves, and test information functions resulting from the use of different IRT software programs were compared. Results from different programs were similar for the simplest IRT models investigated, but differed for the more complex IRT models.

Chapter 5

In Chapter 5, item parameter estimates from Chapter 4 were used to conduct IRT true score equating, to examine how much differences noted in item parameter estimates affected estimated IRT true score equating relationships. Consistent with what was found in Chapter 4, IRT true score equating results from different programs were similar for the simplest IRT models investigated, but differed for the more complex IRT models.

Chapter 6

Chapter 6 investigates the use of various factor analysis-based methods for evaluating test dimensionality of mixed-format tests using data from AP examinations. The use of intact test forms and examinee groups allows for the comparison of general patterns of similarities and

dissimilarities among results from different methods in realistic settings. Focus was on understanding general patterns of similarities and dissimilarities among results from different factor analysis-based methods. Dimensionality assessment is important for IRT applications, including equating, and very little research exists for assessing dimensionality of mixed-format tests.

Chapter 7

In Chapter 7, a bi-factor multidimensional IRT observed-score equating procedure for mixed-format tests was developed. The appropriateness of the procedure was investigated using data from AP examinations. Results from the proposed procedure were compared to those for unidimensional IRT observed score equating. When there was non-negligible multidimensionality, the bi-factor multidimensional IRT method was more adequate than the unidimensional IRT procedure.

Chapter 8

In this chapter the adequacy of various multidimensional IRT, unidimensional IRT, and traditional equating methods were compared using data from the AP examinations. The general finding was that multidimensional IRT equating was more adequate than unidimensional IRT when the tests were clearly multidimensional. All of the methods investigated tended to produce similar results when the tests appeared to be mainly unidimensional.

Discussion and Conclusions

Chapters 2 and 3 investigated the adequacy of equating mixed-format tests using MC-only and representative sets of common items. These chapters followed from work done in Volume 2. The remaining chapters focused on IRT-related methodology. Chapters 4 and 5 examined IRT parameter estimation and its effects on IRT true score equating. Chapter 6 examined methodology for assessing dimensionality. Chapters 7 and 8 investigated multidimensional IRT equating methods.

As a whole, this volume along with Volume 1 and Volume 2 address many of the important psychometric issues associated with equating mixed-format tests. This work also reflects the use of a variety of different approaches to evaluating equating methodology including the use of real and simulated data-based criteria for making these evaluations.

References

Kolen, M. J., & Lee, W. (Eds.). (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1). (CASMA Monograph Number 2.1)*. Iowa City, IA: CASMA, The University of Iowa.

Kolen, M. J., & Lee, W. (Eds.). (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2). (CASMA Monograph Number 2.2)*. Iowa City, IA: CASMA, The University of Iowa.

Lee, W., He, Y., Hagge, S. L., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2). (CASMA Monograph Number 2.2)* (pp. 13-44). Iowa City, IA: CASMA, The University of Iowa.

Chapter 2: An Investigation of Performance of Equating for Mixed-Format Tests Using Only Multiple-Choice Common Items

Seohong Pak and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

Inclusion of free-response (FR) items in an anchor test (i.e., a common-item set) has some practical limitations, which often results in anchor tests having multiple-choice (MC) items only. Using a series of simulation conditions, the present study evaluates the feasibility of equating mixed-format tests with limited common items (i.e., including MC items only). Some major findings of this study were: (a) a large differential level of MC and FR section correlation between the old and new form groups produced equating results with an unacceptable level of bias; (b) a large effect size for the FR section in the new group caused a large amount of equating bias; (c) when the effect sizes for the MC and FR sections were different, the frequency estimation methods tended to perform better than the chained equipercentile methods; and (d) equating results in terms of AP grades were found to be adequate across most of the simulation conditions.

An Investigation of Performance of Equating for Mixed-Format Tests Using Only Multiple-Choice Common Items

Multiple-choice (MC) items and free-response (FR) items are two of the most frequently used item types in real-world testing situations. These two item types often are mixed in a single test form together, with a belief that they measure different but correlated constructs in the test domain (Hendrickson, Patterson, & Ewing, 2011). However, in spite of benefits of FR items (e.g., measuring higher-order reasoning skills), including FR items in a common-item (CI) set in equating may be vulnerable to issues such as item security, reliability, and changes in rater leniency. Because of these issues, a CI set in alternate forms of a mixed-format test typically consists only of MC items.

Since mixed-format tests have become more popular, several studies have been done to investigate the effect of using only MC items in a CI set for linking or equating with mixed-format tests (He, 2011; Kim & Kolen, 2006; Kim, Walker, & McHale, 2008, 2010; Lee, He, Hagge, Wang, & Kolen, 2012; Liu & Kolen, 2011; Tan, Kim, Paek, & Xiang, 2009; Tate, 2003; Wang, 2013). These studies have found that using an MC-only CI set may lead to biased equating results, because an MC-only CI set violates a prerequisite condition of its representativeness to the total test in equating. Furthermore, several previous studies investigated various factors that could affect results for mixed-format test equating with an MC-only CI set. These studies showed that the correlation between the scores on the CI set and the test as a whole, and group ability differences were important factors.

The present study is an extension of the work by Lee et al. (2012), which evaluated the feasibility of equating mixed-format tests having an MC-only CI set (anchor test) using simulation. In particular, Lee et al. (2012) investigated the effects on equating of (a) the correlation between the two constructs measured by MC and FR items, and (b) group ability differences between the new and old form groups. Their results showed that more error was associated with a lower correlation level and a larger effect size. They also found that as the effect size increased, a higher level of correlation was needed to obtain acceptable equating results. They identified a degree of correlation and an effect size necessary to obtain adequate equating results across different conditions.

Recognizing the fact that Lee et al. (2012) used a limited set of simulation conditions, the present study is intended to extend their work by incorporating more conditions that were not

previously considered. Table 1 summarizes the differences in conditions between the two studies. By incorporating more conditions, the present study investigates (1) how overall equating results compare across various combinations of conditions including (a) sample sizes, (b) anchor test lengths, (c) group ability differences (effect sizes), (d) correlations, (e) equating methods, and (f) smoothing methods; and (2) minimal conditions to get acceptable equating results for a mixed-format test with a CI set consisting solely of MC items. A series of simulations was conducted to address these issues.

Method

Data from two forms (old and new) of the AP Spanish Literature exam were used as the basis for simulation. Since Lee et al. (2012) used the AP World History exam, some of the results from the present study might not be directly comparable with those from Lee et al. (2012). Each form of the Spanish Literature exam was comprised of 65 MC items and 7 FR items. Among the 65 MC items, 23 MC items were used as an anchor test. The sample sizes for the old and new form data were 1,087 and 1,943, respectively, after eliminating all the examinee records with missing item responses. For this simulation study, the three parameter logistic model (Lord, 1980) was used to fit the MC items and the graded response model (Samejima, 1997) was used for fitting the FR items.

Two sets of item parameter estimates for the anchor test were yielded because calibration was done separately for each form using PARSCALE (Muraki & Bock, 2003). Once these two sets were obtained, the parameter estimates for the anchor test in the new form were replaced with those in the old form. The estimated item parameters were treated as true item parameters for each of two forms to simulate data. To establish the true equating relationships, simple-structure multidimensional IRT observed-score equating (Lee & Brossman, 2012) was conducted based on the population ability distributions and true item parameters. The simple-structure multidimensional model was used as a framework for simulating data, and thereby was used to establish criterion equating relationships. When simulation conditions varied in bivariate population distributions for the old and new form groups, different true equating relationships were determined. The general procedures for simulation used in this study were similar to Lee et al.'s (2012) study.

Simulation Factors

A total of 486 ($3 \times 9 \times 9 \times 2$) conditions were included in the present study. As shown in Table 1, simulation conditions were:

- three different sample sizes,
- nine differential levels of effect size differences between the old and new form groups,
- nine differential levels of correlations between the old and new form groups, and
- two different anchor test lengths.

Note that the population means (effect sizes) for the MC and FR items in the old form group were always set to zero, so that the population taking the new form was always more able. The results were obtained for each of twelve traditional equating methods and each of three types of score scales. The traditional equating methods used here were the frequency estimation (FE) method and the chained equipercentile (CH) method (Kolen & Brennan, 2014). Each of these two methods was used with the following smoothing methods: (a) no smoothing, (b) log-linear presmoothing with smoothing parameters of either 4 or 6 for the marginal and 1 for the cross product (i.e., 4, 4, & 1 and 6, 6, & 1), and (c) cubic-spline postsmoothing with smoothing parameters of .1, .3, and .5. The three types of score scales used in equating were: raw scores (summed composite scores of MC and FR items) ranging from 0 to 108, normalized scale scores (NSS) ranging from 0-70, and AP grades ranging from 1-5. Equating results were evaluated by comparing the new-form true and estimated equating relationships in terms of unrounded equated scores for all types of score scales.

Evaluation Indices

Three indices were used for evaluating the equating results for twelve equating methods over 100 replications: mean squared error (MSE), squared bias (SB), and variance (VAR). These indices were computed for each raw-score point and across all score points. The conditional SB was computed as:

$$\mathrm{SB}(x) = \left[\left(\frac{1}{100}\sum_{r=1}^{100}\hat{e}_{xr}\right) - e_x\right]^2, \quad (1)$$

where e_x is the true equated score at raw score x, and $\hat{e}_{xr}$ is an estimated equated score at raw score x on replication r. That is, the conditional SB is defined as the square of the difference

between a mean equated score over 100 replications and a true equated score on a particular raw score point. The conditional variance was computed by

$$\mathrm{VAR}(x) = \frac{1}{100}\sum_{r=1}^{100}\left[\hat{e}_{xr} - \left(\frac{1}{100}\sum_{r=1}^{100}\hat{e}_{xr}\right)\right]^2. \quad (2)$$

The conditional MSE was computed as the sum of the conditional SB and conditional variance. The corresponding overall summary statistics were computed as the average of conditional summary statistics across all score points, weighted by the frequencies of new-form scores on the new-form population computed based on the item response theory (IRT) model used to simulate the data.

In addition to these three indices, the Difference That Matters (DTM) was used to determine an acceptable level of equating error. The DTM was defined as .5 in the present study according to the rationale suggested by Dorans and Feigenbaum (1994). The squared DTM value of .25 was compared to $\mathrm{SB}(x)$. When $\mathrm{SB}(x)$ was greater than .25, the equating results were considered unacceptable.

Results

Main Effects

In this main effects section, results for each study factor are discussed. The overall summary indices reported in this section are those that are aggregated (i.e., averaged) over all other conditions except for the study factor under consideration.

Sample size and anchor test length. As presented in Table 2, MSE decreased as the sample size increased, mainly due to a decrease in VAR. Note that the magnitudes of SB were very similar between the 3,000 and 9,000 conditions. The aggregated overall statistics for the full-anchor-test-length condition were always smaller than those for the half-anchor-test-length condition. These findings for sample size and anchor test length were expected, as it is always anticipated that equating is improved if larger sample size and/or longer anchor test length is used. The results for all three types of score scales showed basically the same pattern.

Correlation. Results for various differential levels of correlation are provided in Table 3. The levels of differential correlation are shown in parentheses with two numbers separated by a comma. The number before a comma indicates the correlation between the MC and FR sections for the old form, while the number after a comma represents the correlation between the MC and FR sections for the new form.

As shown in Table 3, the overall aggregated statistics for the main effect of correlation showed somewhat different patterns across the three types of score scales. For raw scores, the (.5 & .95) condition yielded the largest MSE, followed by (.95 & .5), (.5 & .8), and (.8 & .5). This sequence was also shown in AP grades. However, for scale scores, the largest MSE was associated with the (.95 & .5) condition, followed by (.5 & .95), (.8 & .5), and (.5 & .8). For all three types of score scales, in general, the smallest values of MSE tended to be observed for the conditions of (.95 & .95), (.8 & .8), and (.5 & .5). The differences in MSE across various correlational conditions were mainly due to differences in SB rather than VAR.

It appeared that a larger difference in the correlation between the MC and FR abilities for the two forms was associated with larger bias. As the difference in the correlation became smaller, the amounts of MSE and SB also decreased. No correlation difference in the two forms generally corresponded to smaller MSE and SB, even with a very low correlation level (i.e., .5). However, for AP grades, the correlation condition of (.5 & .5) was not better than the conditions of (.8 & .95) and (.95 & .8) in terms of MSE.

Effect size. The aggregated overall statistics for the main effect of effect size are provided in Table 4. Various levels of effect size are indicated in parentheses in Table 4. The number before a comma indicates the effect size for the MC section, and the number after a comma is the effect size for the FR section.

As shown in Table 4, results for the condition of (.0 & .3) always showed the largest MSE, while results for the (.0 & .0) condition had the smallest MSE across all score types. The second largest MSE condition, (.1 & .3), and the third largest MSE condition, (.3 & .3), were also identical for all score types. The relatively larger MSE values for these conditions came from the relatively larger SB values, but not necessarily the VAR values. Note that all these three conditions had the largest effect size for the FR section (i.e., .3). This suggests that a large effect size in the FR section is a major source of equating error when equating is conducted through MC common items, which fails to separate the group difference and item-difficulty difference in the FR section.

Equating methods. As shown in Table 5, each of the MSE, SB, and VAR statistics was aggregated for the main effect of two traditional equating methods (FE and CH). The shaded cells indicate the smaller value of each overall statistic between the FE and CH methods. All three types of score scales yielded the same pattern, in terms of MSE and SB, in which the CH

method had smaller amount of error than the FE method. However, VAR was higher for the CH method than the FE method. All six FE equating methods always yielded higher MSE and SB, while all six CH equating methods had higher VAR. These patterns were the same across all score scales.

Smoothing methods. The unsmoothed, presmoothing, and postsmoothing methods were compared in terms of the aggregated overall error as presented in Table 6. Again, the shaded cells show the smallest value of each statistic among the three smoothing methods. For raw scores, the overall SB was almost identical across three types of smoothing methods. However, the overall VAR was different — the largest VAR was shown in the unsmoothed condition, while the smallest amount was found with the presmoothing method. These differences had a direct impact on the overall MSE — the largest MSE with the unsmoothed and the least MSE with the presmoothing method.

In terms of the overall MSE and VAR statistics, scale scores had the same pattern as raw scores. However, the smallest overall SB was found with the presmoothing method. In addition, the unsmoothed condition had slightly smaller overall SB than the postsmoothing method. For AP grades, the presmoothing method had the largest overall SB and the smallest overall VAR. As for the other two score scales, the unsmoothed method showed the largest overall MSE. However, unlike the other score scales, the postsmoothing method was associated with the smallest MSE.

Main Effect Summary

Clearly, larger sample size or longer anchor test length helped reduce MSE. As either the correlation between the MC and FR abilities increased or the correlation for the two forms became more similar, MSE decreased. MSE increased as the effect size for the MC and FR sections increased. As the effect size for the FR section became larger, MSE increased much more than the conditions with larger effect size for the MC section. The CH method was better in terms of having smaller amount of aggregated MSE. As expected, the unsmoothed condition had the largest MSE. Among the six main effects, sample size, anchor test length, and equating methods yielded the same pattern across all score scale types. By contrast, for the main effects of correlation, effect size, and smoothing methods showed different patterns across score scales.

Interaction Effects

The interaction effects are also evaluated using the aggregated overall MSE, SB, and VAR. The interactions between: (1) correlations and equating methods, (2) correlations and smoothing methods, (3) effect sizes and equating methods, and (4) effect sizes and smoothing methods, are mainly described.

Correlations and equating methods. As shown in Table 7, all score scale types showed smaller overall VAR for the FE method across all correlation conditions. However, the three score scale types had different patterns in terms of the MSE and SB statistics.

For raw scores, MSE for CH was smaller than for FE for all conditions, except for one condition: correlation of .5 for the old form, and .95 for the new form. Also, relatively larger differences between the two equating methods were observed, when the new form had a correlation of .5. In terms of SB, the CH method had smaller bias for all correlation conditions; however, the CH method yielded larger VAR for all correlation conditions. Across various differential correlation conditions, the relative performance of the two equating methods was very similar. Likewise, the pattern of magnitudes of error across correlation conditions was similar for the two equating methods.

Scale scores showed an irregular pattern. When the new form had a correlation of .5, the CH method performed better showing smaller MSE and SB. By contrast, when the old form had a low correlation of .5 and the new form had different correlation levels of .8 or .95, the FE method had smaller MSE and SB. Other than these, there was no predictable pattern according to the various correlation levels.

No interaction effect was found for AP grades. That is, regardless of the correlation conditions, the FE method always had larger MSE and SB, and smaller VAR. The performance of the two equating methods were also similar across correlation conditions — larger MSE and SB when the correlation differences between the new-form and old-form groups were large.

In sum, there seemed no strong interaction effect between the differential correlation levels and equating methods. Among the three score scales, scale scores showed a somewhat larger interaction effect than the other two score scale types — i.e., four conditions showed some interaction effects for scale scores, while only one condition for raw scores and no condition for AP grades showed interaction effects.

Correlations and smoothing methods. As presented in Table 8, all three score scale types showed smaller VAR for the presmoothing method across all correlation levels. However, different patterns in terms of MSE and SB were found for different score scales.

For raw scores, the postsmoothing method showed slightly lower MSE for three correlation conditions — i.e., (.8 & .5), (.95 & .5), and (.95 & .8), while the presmoothing method had slightly lower MSE for the other six conditions. The pattern of the interaction effect appeared mixed. In particular, when two forms had the same level of correlation, the unsmoothed method was associated with the smallest SB. The presmoothing method showed the smallest VAR for all conditions.

For scale scores, MSE did not have any interaction effect with the presmoothing method showing the lowest MSE across all conditions. There were three conditions where either the postsmoothing method or unsmoothed method showed smaller SB than the presmoothing method — i.e., (.5 & .5), (.5 & .8), and (.8 & .95).

For AP grades, the postsmoothing method was associated with the lowest MSE, while the presmoothing method had the lowest VAR across all correlation conditions. The unsmoothed and postsmoothing methods showed very similar SB that was smaller than SB for the presmoothing method.

The results for various differential correlation levels were consistent for three types of smoothing methods and three types of score scales. Large differences in the correlation of the MC and FR abilities between the old and new forms were always associated with large MSE and SB.

In sum, the main effects of the smoothing methods and differential correlation levels were mainly maintained in this interaction analysis, except that there was an irregular interaction effect between the smoothing methods and correlation levels in SB for raw scores.

Effect sizes and equating methods. As shown in Table 9, the interaction between effect size and equating methods showed the same pattern for MSE and SB for all score scales. The CH method showed lower MSE and SB than the FE method for the following three correlation conditions for all score scales: (.1 & .1), (.1 & .3), and (.3 & .3). The FE method had lower MSE and SB than the CH method for the other six correlation conditions. However, there was no interaction for VAR — the FE method always had lower VAR than CH for all correlation conditions for all score scales. The lower MSE and SB for the CH method under only three

correlation conditions out of nine seemed inconsistent with the main effect that supported, overall, the better performance of CH than FE. The overall lower error for CH appeared to be due to the substantially smaller MSE for CH relative to MSE for FE under the (.3 & .3) condition. This example clearly suggests that the main-effect results are overly simplified ones, albeit useful in making overall comparisons, and should be used along with other information such as interaction effects involving other important factors. Note also that, regardless of the score scale types, the difference between the two equating methods became larger in the following conditions: (.1 & .3), (.3 & .0), and (.3 & .3).

Effect sizes and smoothing methods. The interaction effects for effect size and smoothing methods are provided in Table 10. For all score scale types, there was no interaction effect in VAR showing the smallest VAR for the presmoothing method. However, each score scale type had different patterns in terms of MSE and SB.

For raw scores, the interaction effect between effect size and the smoothing methods in the MSE showed a pattern where the postsmoothing method had lower MSE than the presmoothing method under three effect-size conditions with the FR effect size of .3. In terms of SB, the interaction effect showed a mixed pattern.

For scale scores, the presmoothing method showed lower MSE and SB across all effect-size conditions, except for (.3 & .3).

For AP grades, the postsmoothing method exhibited the lowest MSE for all effect-size conditions, except for two conditions (.1 & .0) and (.3 & .0), for which the presmoothing method had lower MSE. The results for SB showed an interesting pattern where the unsmoothed method had the lowest MSE for four effect-size conditions including (.0 & .0), (.1 & .0), (.3 & .0), and (.3 & .1).

The pattern of results for various effect-size conditions tended to be similar for all smoothing methods and types of score scales and consistent with the main effect of effect size. MSE and SB were larger for the effect-size conditions with the FR effect size of .3.

Difference That Matters (DTM)

In order to determine an acceptable level of equating bias, the SB values were compared to the squared DTM value of .25. SB was presumed to be more appropriate than MSE when it comes to the comparison with a DTM in that both the SB and DTM are concerned about "differences" in equivalents. Each of twelve equating methods was compared to the DTM of .25

for each combination of simulation factors. The DTM results are presented in two tables. Table 11 provides results for equal correlation conditions, and Tables 12 and 13 summarize results for differential correlation conditions. In each table, the following letter/number identifiers were used to indicate which equating methods produced adequate equating results for each combination of conditions:

- A: all twelve equating methods had SB smaller than the DTM;
- N: none of twelve equating methods had SB smaller than the DTM;
- C: all six CH methods had SB smaller than the DTM;
- F: all six FE methods had SB smaller than the DTM;
- R: all six presmoothing methods had SB smaller than the DTM;
- S: all six postsmoothing methods had SB smaller than the DTM;
- Rc: all presmoothing with the CH methods had SB smaller than the DTM;
- Rf: all presmoothing with the FE methods had SB smaller than the DTM;
- Sc: all postsmoothing with the CH methods had SB smaller than the DTM;
- Sf: all postsmoothing with the FE methods had SB smaller than the DTM;
- +n: only 'n' number of equating methods met the DTM criterion; and
- −n: all but 'n' number of equating methods met the DTM criterion.

Note that the cells that are not highlighted (i.e., "N") represent conditions where no equating method produced adequate equating results with bias below the DTM level.

Raw scores. When the correlation levels for the two forms were identical, the magnitude of SB seemed mainly affected by effect size, but not influenced much by either sample size or anchor test length, as shown in Table 11. Specifically, when the effect size for the FR section was zero, the combined condition comprised of the smallest sample size, the lowest correlation levels for both forms, and the half length of anchor test was enough to get adequate equating results for all equating methods. However, when there existed any amount of effect size for both the MC and FR sections, the results depended on sample size, correlation, and anchor test length. For example, for the effect size condition of (.1 & .1), larger sample size, the same but higher correlation levels, or the full length of anchor test performed better in terms of getting less than .25 of squared bias. It is interesting to note that when the effect size for the MC section is large and there is no effect size for the FR section (i.e., (.3 & .0)), there are many conditions where only the results for the FE methods were acceptable. By contrast, for the effect size

conditions of (.1 & .1) and (.3 & .3), there were conditions where only the CE methods produced acceptable equating results. Note that in Lee et al. (2012) study, the CH methods always performed better than the FE methods, but these results were limited to the conditions of the same effect sizes for both the MC and FR sections. However, in the present study, when the effect sizes for both sections were different, the FE methods sometimes showed smaller bias than did the CH method.

As shown in Table 12, when there was some difference between the correlation levels for the two forms, none of equating results had an acceptable level of bias. Only a few equating methods were acceptable for a very high correlation condition of (.95 & .8) as presented in Table 13. Furthermore, for the conditions with differential correlation levels, neither sample size nor anchor test length seemed to help getting acceptable equating results.

Scale scores. When the correlations of the two forms were in the same level, length of the anchor test seemed to have little effect on reducing equating bias as shown in Table 11. However, for the condition of the largest effect sizes for both sections, (.3 & .3), the full length anchor test performed better with the CH methods being acceptable. As for the raw-score results, when the effect size for the FR section was .3 and the effect size for the MC section was smaller than .3, there were no equating methods that produced acceptable equating results for any combination of conditions. The conditions of little or no effect size for the MC and FR sections seemed to be less affected by sample size and anchor test length.

On the other hand, when the correlations of the two forms were different (i.e., Table 12), no equating methods were found to have an acceptable magnitude of equating bias when the difference in correlation is large between the two forms. However, when the correlations for both forms were large and the differences in the correlation levels were small (i.e., Table 13), some equating methods showed SB smaller than .25 under some limited effect-size conditions — i.e., the effect sizes for the MC and FR sections were zero or .1; or the effect size for the FR section was smaller than .3. Under the differential correlation levels, sample size and anchor test length tended to be effective in reducing the amount of SB.

AP grades. As shown in Tables 11 through 13, for AP grades, all equating methods produced results with SB that was always lower than the DTM criterion, regardless of the correlation levels.

Summary and Discussion

Investigating the effects of various factors and finding proper equating methods under particular conditions have been difficult problems to address, as the format of tests has become more diverse and more complicated. One popular test format consists of both MC and FR items, often referred to as a mixed-format test. Mixed-format tests have several issues in conducting equating. For example, it is difficult for a common-item set in a mixed-format test to be representative of the entire test in the sense that the common items contain all item types that are in the full length test form. Due to practical constraints, common items in a mixed-format test usually are comprised only of MC items. This results in the violation of the mini-version property of the common-item set in equating, so it is worth investigating several factors that affect the results of equating for mixed format tests with only MC items as a common-item set. The present study is an extended effort to develop some guidelines in selecting equating methods for mixed-format tests under various simulation conditions.

The results of this study show that the magnitude of error decreases as sample size increases. The amount of error also decreases as the length of anchor test (a common-item set) increases. In general, the error increases as effect sizes of the group ability differences for both sections increase. Besides, as the correlation between the MC and FR sections decreases, the magnitude of error increases. These results are consistent with findings from previous studies (Cao, 2008; Hagge & Kolen, 2011; Kirkpatrick, 2005; Lee et al., 2012; Walker & Kim, 2009).

More specifically, when all other factors are controlled to be the same, the main effect of effect size becomes obvious — i.e., the condition of (.0 & .3) produces the largest amount of aggregated error across all score scale types. Overall, when effect size for the FR section is large, more error seems to be introduced. In addition, in terms of investigating the main effect of correlation between the MC and FR abilities, large differences in correlation between the two forms increase the amount of error. This pattern is observable for all score scale types. In addition, in terms of interaction effects, even *no* group ability difference cannot help reduce the amount of error when there exists some difference in correlation between the two forms. A large sample size is not helpful in reducing the amount of error either, when a large difference in correlation between the two forms exists. In other words, the effect of correlation differences between two forms is too large to reduce the amount error even with a large sample size and no group ability differences in the MC and FR sections.

When considering the main effect of the equating methods, all score scale types show less aggregated MSE and SB for the CH method. And this finding is consistent with the results reported in Lee et al. (2012) and Wang, Lee, Brennan, and Kolen (2008). In comparing smoothing methods, both the presmoothing and postsmoothing methods help reduce aggregated VAR across all score scale types. For raw scores and AP grades, the postsmoothing methods perform better in terms of less amount of aggregated SB, while the presmoothing performs better for scale scores.

In terms of the DTM criterion, it is conclusive that the correlation between two forms should be the same or as similar as possible to get an acceptable level of equating bias. Also, conditions of small or no effect size for both MC and FR sections are necessary for raw scores and scale scores. However, a large effect size of the MC section seems to less influence on the magnitude of equating bias. For scale scores, the correlation levels, which yield an acceptable magnitude of equating bias are not as strict as raw scores — i.e., the correlations for the two forms need not be the same. On the other hand, no restriction is required in the correlations for AP grades to get an acceptable magnitude of equating bias.

MSE and the SB seem to be affected profoundly by two conditions across all score scale types: (a) the correlation between the MC and FR sections, and (b) the effect sizes for the MC and FR sections in the new group. On the other hand, VAR seems to be easily affected by the other two conditions: (a) sample size, and (b) length of anchor test. This means that the magnitude of VAR shows relatively stable patterns across various effect-size and correlation conditions, if sample size and anchor test length are controlled to be the same.

Overall, even with a large sample size and full anchor test length, the amount of error in equating gets larger as the difference in correlations between two forms becomes larger and the effect size gets larger. The minimal group ability difference required for adequate equating for a mixed-format test forms seems to depend largely on the correlation conditions.

From the results of this study — i.e., using the DTM criterion, the minimal condition to get an acceptable level of equating bias could be identified. When there is no effect size for both the MC and FR sections or little effect size exists only for the MC section, the lowest correlation level examined in this study (i.e., .5) for both forms is acceptable. Under these effect-size conditions, the half-length of a CI set with small sample size (i.e., 1,000) is enough. This result is applicable only when the correlations for both forms are the same with having all three other

minimum conditions (the correlation, the anchor test length, and the sample size). This finding is the same for all score scale types.

When there is little effect size for both MC and FR sections, the minimal correlation level needed for all twelve equating methods to have SB less than .25 is .8 for both forms. The sample size of 1,000 is sufficient. However, the minimal anchor test length condition is different — i.e., for raw scores, the full length of anchor test is needed, while the half length of anchor test is enough for scale scores and AP grades.

It is relatively hard to find the minimal combination conditions when the effect sizes for both sections become larger. If there is a large effect size of the MC section and little effect size for the FR section, a correlation level of .5 is sufficient with a full length anchor test and sample size of 1,000, or a half-length anchor test combined with sample size of 3,000 for raw scores. For scale scores, the correlation level of .5 with a full length anchor test, or .8 with a half-length anchor test is acceptable. The sample size of 1,000 is enough. If there is a large effect size for both sections, the minimal correlation level needed is .95 for the raw scores and .8 for scale scores. Both score scale types need a full length of anchor test with a small sample size of 1,000. However, for this particular condition, only CH methods can have an acceptable level of equating bias.

There are some limitations in this study. First, even though nine pairs of effect size were used, it is still possible to consider a wider range of effect sizes. Since the influence of the effect size is large for raw scores and scale scores, a smaller unit of effect size is worth investigating. Second, this study assumes that the simple-structure multidimensional IRT model for the true equating relationship holds in the simulation, which might affect the results of the traditional equating methods that are based on the different assumptions. It would be interesting to replicate this study using different equating criteria and/or different simulation designs to see if the same conclusions are drawn. Last, the mixed-format test used in this study only consisted of two types of items with certain characteristics such as the number of MC and FR items, difficulty levels, etc. Therefore, generalizing the results of this study should be done with caution. In spite of these limitations, this study aims to provide practitioners with useful guidelines when it comes to equating mixed-format tests.

References

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets.* Unpublished doctoral study, University of Maryland.

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

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1).* (*CASMA Monograph No. 2.1*) (pp. 95-136). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

He, Y. (2011). *Evaluating equating properties for mixed-format tests*. Unpublished doctoral study, University of Iowa.

Hendrickson, A., Patterson, B., & Ewing, M. (2011, May). *Developing form assembly specifications for exams with multiple choice and constructed response items*. Paper presented at the 2011 annual conference of the National Council on Measurement in Education, Denver, CO.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*, 357-381.

Kim, S., Walker, M. E., & McHale, F. (2008). *Equating of mixed-format tests in large-scale assessments.* Technical Report (RR-08-26). Princeton, NJ: Educational Testing Service.

Kim, S., Walker, M. E., & McHale, F. (2010). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement, 47*, 186-201.

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating.* Unpublished doctoral study, University of Iowa.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-formant tests using a simple-structure multidimensional IRT framework. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*) (pp. 115-142). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed format tests using dichotomous common items. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*) (pp. 13-44). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1).* (*CASMA Monograph No. 2.1*) (pp. 75-94). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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

Muraki, E., & Bock, R. (2003). PARSCALE (Version 4.1) [Computer program]. Chicago, IL: Scientific Software International.

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. K. Hambleton (Ed.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer-Verlag.

Tan, X., Kim, S., Paek, I., & Xiang, B. (2009). *An alternative to the trend scoring method for adjusting scoring shifts in mixed-format tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Tate, R. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement, 63*, 893-914.

Walker, M. E., & Kim, S. (2009, April). *Linking mixed-format tests using multiple choice anchors*. Paper presented at the 2009 annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*, 632-651.

Wang, W. (2013). *Mixed-format test score equating: effect of item-type multidimensionality, length and composition of common-item set, and group ability difference.* Unpublished doctoral study, University of Iowa.

Table 1

Comparison of Simulation Conditions Between Lee et al. (2012) and the Present Study

<table>
<tr><th>Simulation Conditions</th><th>Lee et al. (2012)</th><th>Present Study</th></tr>
<tr><td>Test Subject</td><td>♦ AP World History</td><td>♦ AP Spanish Literature</td></tr>
<tr><td>Sample Sizes</td><td>♦ One Sample Size (3,000)</td><td>♦ Three Sample Sizes (1,000/ 3,000/ 9,000)</td></tr>
<tr><td>Equating Methods</td><td>♦ 6 traditional equating methods
· Unsmoothed FE & CH
· Log-linear presmoothed FE & CH (using 6, 6, and 1)
· Cubic-spline postsmoothed FE & CH (using S=.1)</td><td>♦ 12 traditional equating methods
· Unsmoothed FE & CH
· Log-linear presmoothed FE & CH (Using 6, 6, and 1 & 4, 4, and 1)
· Cubic-spline postsmoothed FE & CH (Using S=.1, S=.3, and S=.5)</td></tr>
<tr><td>Effect Sizes</td><td>♦ 5 levels of effect size (.05, .1, .2, .3, or .5)
<table>
<tr><th colspan="2">Old Group</th><th colspan="2">New Group</th></tr>
<tr><th>MC</th><th>FR</th><th>MC</th><th>FR</th></tr>
<tr><td>.0</td><td>.0</td><td>.05</td><td>.05</td></tr>
<tr><td>.0</td><td>.0</td><td>.1</td><td>.1</td></tr>
<tr><td>.0</td><td>.0</td><td>.2</td><td>.2</td></tr>
<tr><td>.0</td><td>.0</td><td>.3</td><td>.3</td></tr>
<tr><td>.0</td><td>.0</td><td>.5</td><td>.5</td></tr>
</table></td><td>♦ 9 differential levels of effect size
<table>
<tr><th colspan="2">Old Group</th><th colspan="2">New Group</th></tr>
<tr><th>MC</th><th>FR</th><th>MC</th><th>FR</th></tr>
<tr><td>.0</td><td>.0</td><td>.0</td><td>.0</td></tr>
<tr><td>.0</td><td>.0</td><td>.0</td><td>.1</td></tr>
<tr><td>.0</td><td>.0</td><td>.0</td><td>.3</td></tr>
<tr><td>.0</td><td>.0</td><td>.1</td><td>.0</td></tr>
<tr><td>.0</td><td>.0</td><td>.1</td><td>.1</td></tr>
<tr><td>.0</td><td>.0</td><td>.1</td><td>.3</td></tr>
<tr><td>.0</td><td>.0</td><td>.3</td><td>.0</td></tr>
<tr><td>.0</td><td>.0</td><td>.3</td><td>.1</td></tr>
<tr><td>.0</td><td>.0</td><td>.3</td><td>.3</td></tr>
</table></td></tr>
<tr><td>Correlations</td><td>♦ 11 levels of correlation between MC and FR section (.5 to 1.0 an increment of .05)</td><td>♦ 9 differential levels of correlation between MC and FR section
<table>
<tr><th>Old Form</th><th>New Form</th></tr>
<tr><td>.5</td><td>.5</td></tr>
<tr><td>.5</td><td>.8</td></tr>
<tr><td>.5</td><td>.95</td></tr>
<tr><td>.8</td><td>.5</td></tr>
<tr><td>.8</td><td>.8</td></tr>
<tr><td>.8</td><td>.95</td></tr>
<tr><td>.95</td><td>.5</td></tr>
<tr><td>.95</td><td>.8</td></tr>
<tr><td>.95</td><td>.95</td></tr>
</table></td></tr>
<tr><td>Anchor Test Length</td><td>♦ 31% of MC Section</td><td>♦ 35% and 18% of MC Section</td></tr>
<tr><td>Score Scales</td><td>♦ Raw, NSS, & AP Grades</td><td>♦ Raw, NSS, & AP Grades</td></tr>
<tr><td>IRT Models</td><td>♦ 3PL + GRM</td><td>♦ 3PL + GRM</td></tr>
</table>

Table 2

Aggregated Overall Statistics for Sample Size and Anchor Test Length

		Sample Size			Anchor Test Length	
		1,000	3,000	9,000	Half	Full
Raw Scores	MSE	2.182	1.930	1.832	2.080	1.883
	SB	1.739	1.771	1.777	1.829	1.696
	VAR	0.443	0.159	0.055	0.251	0.187
Scale Scores	MSE	1.414	1.288	1.221	1.356	1.259
	SB	1.145	1.186	1.186	1.203	1.141
	VAR	0.270	0.102	0.035	0.153	0.118
AP Grades	MSE	0.059	0.053	0.051	0.056	0.053
	SB	0.043	0.046	0.049	0.047	0.045
	VAR	0.016	0.007	0.003	0.010	0.007

Table 3

Aggregated Overall Statistics for Correlations

		Correlation between MC and FR for (old form & new form)								
		(.5 & .5)	(.5 & .8)	(.5 & .95)	(.8 & .5)	(.8& .8)	(.8 & .95)	(.95 & .5)	(.95 & .8)	(.95 & .95)
Raw Scores	MSE	1.127	2.348	3.842	2.245	1.036	1.357	3.593	1.263	1.020
	SB	0.910	2.115	3.603	2.038	0.821	1.134	3.392	1.047	0.799
	VAR	0.218	0.233	0.238	0.208	0.215	0.223	0.201	0.215	0.220
Scale Scores	MSE	0.633	1.224	2.217	1.837	0.600	0.716	3.002	0.937	0.605
	SB	0.501	1.080	2.068	1.712	0.466	0.572	2.880	0.804	0.466
	VAR	0.131	0.143	0.149	0.125	0.134	0.144	0.122	0.133	0.139
AP Grades	MSE	0.047	0.065	0.083	0.059	0.040	0.044	0.074	0.042	0.038
	SB	0.038	0.056	0.074	0.050	0.031	0.036	0.066	0.033	0.030
	VAR	0.009	0.009	0.009	0.009	0.008	0.008	0.009	0.008	0.008

Table 4

Aggregated Overall Statistics for Effect Sizes

		Effect Size (MC & FR) for the New group								
		(.0 & .0)	(.0 & .1)	(.0 & .3)	(.1 & .0)	(.1& .1)	(.1 & .3)	(.3 & .0)	(.3 & .1)	(.3 & .3)
Raw Scores	MSE	1.173	1.471	3.870	1.197	1.313	3.328	1.712	1.325	2.442
	SB	0.948	1.248	3.651	0.981	1.092	3.110	1.494	1.108	2.227
	VAR	0.225	0.223	0.219	0.216	0.221	0.218	0.218	0.217	0.215
Scale Scores	MSE	0.795	0.961	2.253	0.851	0.920	2.032	1.274	1.027	1.660
	SB	0.667	0.834	2.127	0.721	0.787	1.900	1.124	0.878	1.513
	VAR	0.128	0.127	0.126	0.129	0.133	0.132	0.150	0.149	0.146
AP Grades	MSE	0.031	0.042	0.109	0.032	0.035	0.094	0.047	0.035	0.066
	SB	0.022	0.033	0.101	0.023	0.027	0.085	0.039	0.027	0.058
	VAR	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.008

Table 5

Aggregated Overall Statistics for Equating Methods

		Raw Scores			Scale Scores			AP Grades		
		MSE	SB	VAR	MSE	SB	VAR	MSE	SB	VAR
Equating Methods	FE	2.027	1.835	0.192	1.326	1.206	0.120	0.056	0.048	0.008
	CH	1.936	1.690	0.246	1.290	1.138	0.152	0.053	0.044	0.009

Table 6

Aggregated Overall Statistics for Smoothing Methods

		Raw Scores			Scale Scores			AP Grades		
		MSE	SB	VAR	MSE	SB	VAR	MSE	SB	VAR
Smoothing Methods	Un	2.088	1.768	0.320	1.379	1.180	0.199	0.057	0.045	0.012
	Pre	1.952	1.762	0.190	1.268	1.149	0.119	0.056	0.048	0.008
	Post	1.965	1.761	0.205	1.311	1.186	0.126	0.053	0.045	0.008

Table 7

Aggregated Overall Statistics for Interaction of Correlations and Equating Methods

			Correlation between MC and FR for (old form & new form)								
			(.5 & .5)	(.5 & .8)	(.5 & .95)	(.8 & .5)	(.8& .8)	(.8 & .95)	(.95 & .5)	(.95 & .8)	(.95 & .95)
Raw Scores	MSE	FE	1.235	2.371	3.838	2.355	1.070	1.358	3.703	1.293	1.020
		CH	1.019	2.325	3.845	2.135	1.002	1.356	3.484	1.232	1.019
	SB	FE	1.046	2.167	3.630	2.174	0.880	1.162	3.526	1.102	0.826
		CH	0.773	2.063	3.577	1.902	0.763	1.107	3.258	0.993	0.773
	VAR	FE	0.189	0.204	0.208	0.182	0.190	0.196	0.177	0.190	0.195
		CH	0.246	0.262	0.268	0.234	0.239	0.250	0.226	0.240	0.246
Scale Scores	MSE	FE	0.688	1.148	2.091	1.970	0.614	0.673	3.165	0.986	0.599
		CH	0.577	1.299	2.344	1.705	0.587	0.759	2.840	0.888	0.611
	SB	FE	0.574	1.023	1.959	1.860	0.494	0.545	3.057	0.868	0.475
		CH	0.429	1.138	2.177	1.565	0.438	0.599	2.704	0.741	0.456
	VAR	FE	0.114	0.126	0.132	0.109	0.119	0.128	0.107	0.118	0.124
		CH	0.149	0.161	0.167	0.140	0.149	0.160	0.136	0.147	0.154
AP Grades	MSE	FE	0.051	0.067	0.084	0.061	0.041	0.045	0.076	0.042	0.038
		CH	0.043	0.063	0.082	0.057	0.039	0.044	0.073	0.041	0.038
	SB	FE	0.042	0.058	0.076	0.053	0.033	0.037	0.068	0.034	0.030
		CH	0.033	0.054	0.072	0.048	0.030	0.035	0.064	0.032	0.029
	VAR	FE	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
		CH	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.009	0.009

Table 8

Aggregated Overall Statistics for Interaction of Correlations and Smoothing Methods

			Correlation between MC and FR for (old form & new form)								
			(.5 & .5)	(.5 & .8)	(.5 & .95)	(.8 & .5)	(.8& .8)	(.8 & .95)	(.95 & .5)	(.95 & .8)	(.95 & .95)
Raw Scores	MSE	Un	1.222	2.469	3.984	2.337	1.136	1.469	3.691	1.358	1.122
		Pre	1.104	2.300	3.779	2.240	1.008	1.316	3.592	1.244	0.988
		Post	1.111	2.340	3.836	2.219	1.022	1.348	3.562	1.243	1.007
	SB	Un	0.907	2.132	3.638	2.036	0.820	1.141	3.397	1.043	0.798
		Pre	0.913	2.097	3.573	2.057	0.822	1.124	3.416	1.057	0.798
		Post	0.908	2.122	3.612	2.025	0.822	1.139	3.375	1.043	0.801
	VAR	Un	0.315	0.337	0.346	0.301	0.317	0.328	0.294	0.315	0.324
		Pre	0.191	0.203	0.207	0.182	0.185	0.192	0.176	0.187	0.190
		Post	0.203	0.218	0.223	0.194	0.201	0.209	0.187	0.201	0.206
Scale Scores	MSE	Un	0.685	1.327	2.348	1.878	0.663	0.795	3.056	0.990	0.670
		Pre	0.616	1.200	2.170	1.775	0.577	0.698	2.904	0.892	0.577
		Post	0.627	1.205	2.205	1.865	0.595	0.702	3.049	0.949	0.602
	SB	Un	0.492	1.118	2.130	1.695	0.463	0.584	2.876	0.794	0.466
		Pre	0.499	1.074	2.042	1.665	0.460	0.573	2.797	0.775	0.455
		Post	0.506	1.072	2.065	1.750	0.472	0.568	2.938	0.827	0.473
	VAR	Un	0.193	0.209	0.218	0.184	0.200	0.210	0.180	0.196	0.205
		Pre	0.116	0.126	0.129	0.111	0.117	0.125	0.107	0.117	0.121
		Post	0.121	0.133	0.140	0.115	0.124	0.134	0.112	0.122	0.129
AP Grades	MSE	Un	0.049	0.067	0.084	0.061	0.042	0.046	0.076	0.044	0.040
		Pre	0.048	0.066	0.085	0.060	0.040	0.045	0.076	0.042	0.038
		Post	0.046	0.064	0.081	0.058	0.039	0.043	0.073	0.041	0.037
	SB	Un	0.037	0.055	0.072	0.049	0.031	0.035	0.065	0.032	0.029
		Pre	0.039	0.058	0.077	0.052	0.033	0.037	0.068	0.034	0.031
		Post	0.037	0.055	0.072	0.049	0.031	0.035	0.065	0.033	0.029
	VAR	Un	0.012	0.012	0.012	0.012	0.011	0.011	0.011	0.011	0.011
		Pre	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007
		Post	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.008	0.008

Table 9

Aggregated Overall Statistics for Interaction of Effect Sizes and Equating Methods

			Effect Size (MC & FR) for the New group								
			(.0 & .0)	(.0 & .1)	(.0 & .3)	(.1 & .0)	(.1& .1)	(.1 & .3)	(.3 & .0)	(.3 & .1)	(.3 & .3)
Raw Scores	MSE	FE	1.145	1.438	3.844	1.135	1.356	3.582	1.380	1.312	3.052
		CH	1.201	1.505	3.896	1.259	1.271	3.074	2.044	1.338	1.833
	SB	FE	0.948	1.244	3.652	0.946	1.163	3.391	1.188	1.120	2.862
		CH	0.949	1.253	3.651	1.017	1.022	2.830	1.800	1.095	1.592
	VAR	FE	0.197	0.194	0.193	0.189	0.193	0.191	0.192	0.192	0.190
		CH	0.252	0.252	0.245	0.242	0.249	0.245	0.243	0.243	0.240
Scale Scores	MSE	FE	0.782	0.945	2.242	0.812	0.938	2.166	1.057	0.997	1.994
		CH	0.808	0.977	2.264	0.889	0.902	1.898	1.490	1.057	1.325
	SB	FE	0.668	0.833	2.130	0.698	0.822	2.049	0.925	0.866	1.865
		CH	0.666	0.836	2.123	0.744	0.752	1.750	1.323	0.891	1.161
	VAR	FE	0.114	0.112	0.112	0.114	0.116	0.117	0.132	0.131	0.129
		CH	0.143	0.142	0.140	0.144	0.150	0.148	0.167	0.166	0.164
AP Grades	MSE	FE	0.030	0.041	0.109	0.030	0.037	0.101	0.037	0.035	0.083
		CH	0.031	0.042	0.110	0.034	0.034	0.087	0.057	0.036	0.050
	SB	FE	0.022	0.033	0.101	0.022	0.029	0.093	0.029	0.027	0.075
		CH	0.022	0.033	0.100	0.024	0.024	0.077	0.048	0.027	0.041
	VAR	FE	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
		CH	0.009	0.010	0.009	0.009	0.009	0.009	0.009	0.009	0.009

Table 10

Aggregated Overall Statistics for Interaction of Effect Sizes and Smoothing Methods

			Effect Size (MC & FR) for the New group								
			(.0 & .0)	(.0 & .1)	(.0 & .3)	(.1 & .0)	(.1& .1)	(.1 & .3)	(.3 & .0)	(.3 & .1)	(.3 & .3)
Raw Scores	MSE	Un	1.276	1.577	3.984	1.298	1.418	3.437	1.826	1.435	2.539
		Pre	1.144	1.447	3.853	1.162	1.285	3.314	1.647	1.281	2.439
		Post	1.159	1.453	3.843	1.187	1.297	3.302	1.717	1.318	2.412
	SB	Un	0.949	1.253	3.664	0.983	1.095	3.118	1.509	1.117	2.225
		Pre	0.948	1.251	3.663	0.974	1.093	3.124	1.458	1.094	2.253
		Post	0.949	1.245	3.639	0.986	1.091	3.099	1.513	1.114	2.211
	VAR	Un	0.327	0.324	0.320	0.315	0.322	0.319	0.317	0.318	0.315
		Pre	0.196	0.195	0.190	0.187	0.192	0.190	0.189	0.188	0.187
		Post	0.210	0.208	0.204	0.201	0.206	0.203	0.204	0.203	0.201
Scale Scores	MSE	Un	0.862	1.030	2.327	0.923	0.990	2.097	1.370	1.107	1.709
		Pre	0.764	0.920	2.187	0.817	0.882	1.988	1.211	0.984	1.654
		Post	0.794	0.966	2.272	0.849	0.922	2.040	1.283	1.029	1.647
	SB	Un	0.673	0.843	2.140	0.731	0.793	1.902	1.152	0.889	1.494
		Pre	0.651	0.808	2.076	0.704	0.766	1.872	1.081	0.854	1.526
		Post	0.675	0.849	2.156	0.729	0.799	1.918	1.144	0.891	1.511
	VAR	Un	0.188	0.186	0.187	0.192	0.196	0.195	0.217	0.218	0.215
		Pre	0.112	0.112	0.110	0.113	0.117	0.116	0.131	0.130	0.128
		Post	0.119	0.117	0.117	0.119	0.123	0.122	0.139	0.138	0.136
AP Grades	MSE	Un	0.033	0.044	0.111	0.034	0.038	0.095	0.049	0.037	0.068
		Pre	0.031	0.043	0.112	0.031	0.036	0.096	0.046	0.035	0.069
		Post	0.030	0.040	0.107	0.031	0.034	0.092	0.048	0.035	0.064
	SB	Un	0.021	0.032	0.099	0.022	0.026	0.084	0.037	0.026	0.057
		Pre	0.023	0.035	0.104	0.024	0.028	0.088	0.038	0.028	0.061
		Post	0.022	0.031	0.099	0.023	0.026	0.083	0.039	0.027	0.056
	VAR	Un	0.012	0.012	0.012	0.012	0.012	0.012	0.011	0.011	0.011
		Pre	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
		Post	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008

Table 11

Results of Comparing the Squared DTM and the SB for the Same Correlation Levels

Raw Scores	*N*=1,000						*N*=3,000						*N*=9,000					
	(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .1)	C	C	C	A	C	A	C	A	C	A	C	A	N	A	-3	A	-2	A
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	-1	F	F	N	N	N	-1	F	F	N	F	N	-1	F	F	N	F	N
(.3 & .1)	C	A	C	F	F	F	C	A	-2	F	A	F	C	A	C	F	A	F
(.3 & .3)	N	N	N	N	A	C	N	N	N	N	N	C	N	N	N	N	N	C
Scale Scores	*N*=1,000						*N*=3,000						*N*=9,000					
	(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .1)	C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	A	F	F	N	F	N	A	F	F	N	F	N	A	F	F	N	F	N
(.3 & .1)	C	A	A	F	A	F	C	A	A	A	A	F	C	A	A	A	A	F
(.3 & .3)	N	N	N	C	N	C	N	N	N	C	N	C	N	N	N	C	N	C
AP Grades	*N*=1,000						*N*=3,000						*N*=9,000					
	(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)		(.5 & .5)		(.8 & .8)		(.95 & .95)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Table 12

Results of Comparing the Squared DTM and the SB for the Large Differential Correlation Levels

Raw Scores	*N*=1,000								*N*=3,000								*N*=9,000							
	(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Scale Scores	*N*=1,000								*N*=3,000								*N*=9,000							
	(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .1)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .3)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
AP Grades	*N*=1,000								*N*=3,000								*N*=9,000							
	(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)		(.5 & .8)		(.5 & .95)		(.8 & .5)		(.95 & .5)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .0)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .1)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .3)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Table 13

Results of Comparing the Squared DTM and the SB for the Small Differential Correlation Levels

Raw Scores	*N*=1,000				*N*=3,000				*N*=9,000			
	(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	N	N	N	R	N	N	N	R	N	N	N	N
(.0 & .1)	N	N	N	N	N	N	N	N	N	N	N	N
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	N	N	N	+4	N	N	F	N	N	N	F	N
(.1 & .1)	N	N	N	+1	N	N	N	C	N	N	N	+1
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	N	N	F	N	N	N	N	N	N	N	Sf	N
(.3 & .1)	N	N	C	F	N	N	C	+4	N	N	C	+4
(.3 & .3)	N	N	N	N	N	N	N	N	N	N	N	N
Scale Scores	*N*=1,000				*N*=3,000				*N*=9,000			
	(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	-5	A	-5	-3	-4	-2	-3	A	-4	-2
(.0 & .1)	A	A	N	N	-2	A	N	N	A	A	N	N
(.0 & .3)	N	N	N	N	N	N	N	N	N	N	N	N
(.1 & .0)	+2	N	-2	A	N	N	A	A	N	N	A	A
(.1 & .1)	-1	A	N	Rc	A	A	N	N	A	A	N	Rc
(.1 & .3)	N	N	N	N	N	N	N	N	N	N	N	N
(.3 & .0)	N	N	F	Rf	N	N	N	N	N	N	F	+1
(.3 & .1)	+4	N	+5	A	+4	N	C	A	+4	N	C	A
(.3 & .3)	N	C	N	N	N	C	N	N	N	C	N	N
AP Grades	*N*=1,000				*N*=3,000				*N*=9,000			
	(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)		(.8 & .95)		(.95 & .8)	
	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full	Half	Full
(.0 & .0)	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .1)	A	A	A	A	A	A	A	A	A	A	A	A
(.0 & .3)	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .0)	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .1)	A	A	A	A	A	A	A	A	A	A	A	A
(.1 & .3)	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .0)	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .1)	A	A	A	A	A	A	A	A	A	A	A	A
(.3 & .3)	A	A	A	A	A	A	A	A	A	A	A	A

Chapter 3: Comparison of the Use of MC Only and Mixed-Format Common Items in Mixed-Format Test Score Equating

Wei Wang and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

This study examines the performance of two different common-item set designs on mixed-format equating, including MC only and mixed-format. Simulated data are used. Two sets of test forms with differences in test form difficulty were constructed from two item pools that mimic certain properties of operational data. Factors of interest in the study are the correlation between the latent variables that underlie two item formats used on the test, composition of the common-item set, and group ability difference. The general findings of the study are as follows: 1) When the test forms to be equated are multidimensional, the mixed-format common-item set produces more accurate equating results than the common-item set using only one item format. 2) Relatively large and large group ability differences result in inaccurate equating results. 3) The two common-item sets produce "acceptable" equating results when no or small group ability difference exists, regardless of item-type test dimensionality; and 4) When moderate group ability difference exists (e.g., .10 in effect size units), to obtain adequate equating results, the MC-only common-item set can be used with the pre-smoothed chained equipercentile method, and the disattenuated correlation between the MC and CR section can be as low as .50.

Comparison of the Use of MC Only and Mixed-Format Common Items in Mixed-Format Test Score Equating

Introduction

Mixed-format tests have become increasingly popular in large-scale testing programs and state assessment systems, such as most of the College Board's Advanced Placement (AP) examinations (http://www.collegeboard.com/student/testing/ap/about.html), the National Assessment of Educational Progress (NAEP; http://nces.ed.gov/nationsreportcard/), Ohio Achievement Tests (http://ohio3-8.success-ode-state-oh-us.info/), and most of the Praxis Series™ tests (http://www.ets.org/praxis). Mixed-format tests commonly use two item types, dichotomously-scored multiple-choice (MC) items and polytomously-scored constructed-response (CR) items. The strengths of MC items lie in broad content coverage, efficiency of administration, as well as reliable and objective scoring; however, they have been criticized for being affected by examinee guessing and for lacking ability to elicit the full spectrum of cognitive activity valued by educators (Ebel & Frisbie, 1991; Ferrara & DeMauro, 2006). CR items, on the other hand, are often used to measure complex skills, and the use of CR items helps avoid random guessing (Livingston, 2009). However, CR items can lack efficiency, are often expensive to develop, and typically produce less reliable scores when subjectively scored (Downing, 2006). The two item types, MC and CR, have their own advantages and disadvantages. Test developers often use a combination of the two item formats to take advantage of the benefits of each item type while compensating for their weaknesses (Reshetar & Melican, 2010).

In mixed-format tests, due to the potential differences of the different item types, the use of multiple item formats may lead to many concerns in assessing measurement characteristics (Bennett, Rock, & Wang, 1991). Many studies have reported that MC and CR items exhibit differences in constructs measured (e.g., Manhart, 1996; Perkhounkova & Dunbar, 1999; Rauch & Hartig, 2010; Sykes, Hou, Hanson, & Wang, 2002; Thissen, Wainer, & Wang, 1994; Traub, 1993; Wainer, Wang, & Thissen, 1994). Therefore, the use of multiple item formats could make a test multidimensional in terms of constructs measured.

When multiple forms of the same mixed-format test are administered to examinees, equating is often conducted to ensure the scores from different forms comparable and further ensure the fairness of the assessment. The common-item non-equivalent groups (CINEG) design

(sometimes also referred to as non-equivalent anchor test design) is a data collection design used for equating purposes, and has been widely adopted in practice. Compared to other data collection designs (e.g., single group design, random groups design), the CINEG design is flexible in terms of administration (Kolen & Brennan, 2014). In this design, in order to estimate group difference accurately and further disentangle group difference from form difference, it is suggested that the equating set should be a "mini version" of the total test in terms of content and statistical characteristics (Kolen & Brennan, 2014).

In practice, when equating mixed-format tests in the CINEG design, CR items are often not used as common items. The major reasons include there being a limited number of CR items to select as common items, test security, and rater leniency (Hagge, 2010; He, 2011; Muraki, Hombo, & Lee, 2000). Several concerns are associated with the use of MC items alone as common items for mixed-format equating with the CINEG design. One serious concern is, as described earlier, test multidimensionality caused by the use of different item types. When only MC items are used as common items, the constructs measured by the common-items set might not represent the construct of the total test. If two forms of the same test are well developed, the construct assessed over the MC and CR items for the forms to be equated might be the same, in which case it would make sense to consider equating scores on the forms. However, if only MC items are used as common items to equate scores on the two forms, the common-item set might not adequately reflect the construct being assessed. That is, among examinees administered the forms to be equated, scores on the MC common items might not adequately reflect group differences in the construct assessed over the MC and CR items. In this case, using MC items alone as common items might not adequately disentangle group differences from form differences, and it could lead to systematic errors in equating results. These concerns are present whenever the construct measured by the MC items differs from that measured by the CR items, and likely are most serious when examinee groups taking the two forms differ considerably in ability. In addition, in mixed-format test score equating, in order for the common-item set to represent the total test, Kirkpatrick (2005) suggested taking the effect of item format into account. That is, the common-item set should contain all types of items on the test. However, when MC items alone are used as the common items, format representativeness is not satisfied, which might result in inaccurate equating results.

Few studies have been conducted to investigate the effect of the common-item set on mixed-format equating. For example, through simulation studies, Kirkpatrick (2005) studied the effect of inclusion or exclusion of CR common items on mixed-format IRT equating results. Kirkpatrick (2005) reported that conclusions were highly influenced by differential performance of examinees on subcontent areas. Hagge (2010) conducted a pseudo test form analysis to explore which common-item set composition resulted in the least bias in mixed-format equating, and found mixed conclusions. Lee, He, Hagge, Wang, and Kolen (2012) conducted a series of simulation studies to evaluate the feasibility of using only MC items as common items in mixed-format test score equating. Their results demonstrated that the correlation between the MC and CR latent variables and the group ability difference played decisive roles. They reported that when a large group ability difference existed, a higher correlation was required to achieve acceptable equating results. None of the previous research provided comprehensive studies on the comparisons of different types of common-item sets (i.e., MC-only common-item set and mixed common-item set including both MC and CR items) in mixed-format equating, especially for multidimensional mixed-format test score equating. Using simulated data, the main purposes of this study are to explore how the conclusions from comparisons of different types of common-item sets in mixed-format equating vary under different conditions: item-type multidimensionality, group ability difference, and equating method.

Methods

Test Configuration

Test length is not a factor of interest in this study. However, it is a potential factor that might impact mixed-format test equating results. To avoid the influence of test length, in this simulation study, the total number of items on the form was fixed to be 76, 64 of which were MC items and the remaining 12 items were CR items. Among the 12 CR items, the first eight items were scored 0 to 4, and the remaining items had possible scores from 0 to 8. Weights for both MC and CR items were set to 1. To compute the composite score for an examinee, the score of the MC items and the score points of each CR item were summed. The range of composite scores was 0 to 128 with an increment of 1.

Item Pool Generation

One MC item pool and one CR item pool were used to provide items on test forms. The item pools were created by using operational data on test forms from the College Board's AP

examinations including Art History, Biology, Chemistry, Comparative Government and Politics, Environmental Science, Physics B, Spanish Language, Spanish Literature, and World History. For each examination, item response theory (IRT) item parameters were estimated simultaneously using MULTILOG (Thissen, Chen, & Bock, 2003) for each examination. The three parameter logistic (3PL) model was used to calibrate MC items, whereas the graded response (GR) model was used for CR items. In the computer runs, theta was scaled to have a mean of 0 and a standard deviation of 1. The estimated MC item parameters were treated as parameters for the MC items in the MC item pool, whereas the estimated CR item parameters were treated as parameters for the CR items in the CR item pool. The items along with the parameters from different examinations were included in the same MC or CR pool. The item pools were constructed in this way so that the item parameters used for the items in the simulation are realistic.

Factors of Interest

Form difficulty difference (FDD). For equating purposes, old and new forms are needed. Test forms to be equated can have somewhat different difficulties. In this study, two levels of form difficulty difference (FDD) were considered: 0 and .25 in the effect size units for standardized mean score differences. Zero indicates that the old and new test forms have the same difficulty level, and .25 represents a relatively large FDD. When the FDD is .25, for example, if composite scores on both forms have the same standard deviation of 10, then the mean composite scores between the two forms differs by 2.5. When difficulty differences existed between forms, in this simulation, it was assumed that the new form is more difficult than the old form.

As described earlier, both old and new forms were constructed from the item pools. When the FDD was 0, items on the two test forms to be equated had the same item parameters, were identical, and thus were equally difficult.

When the FDD was .25, the old form was the same as the old form used with the condition of no form difficulty difference. From the old form, 32 MC items and 3 CR items (two with 5 categories and one with 9 categories) were selected to be common with the new form. Unique items on the new form were selected from the item pools to satisfy the target FDD of .25.

Although the test forms shared 32 MC items and 3 CR items, it is important to note that only a portion of these items were used as common items in equating, depending on the

conditions used in the simulation. These specific common item conditions are described in a later part of this section. In addition, the common items were the same for both FDD conditions.

Correlation between MC and CR sections (MC-CR COR). Item-type multidimensionality was considered in this study. The level of multidimensionality was manipulated by controlling the correlation between the two constructs measured by MC and CR items (hereafter, this correlation is abbreviated as MC-CR COR), which was expressed in the IRT theta metric. Four levels of MC-CR COR were considered: .50, .75, .90, and 1.0. The levels of MC-CR COR were selected according to estimated classical disattenuated correlations of the MC and CR scores for the large-scale assessments used for building the item pools, which typically ranged from .75 to 1.0. .50 was selected as the lowest MC-CR COR in order to establish a wide range of correlations although it is extremely low and would not likely occur in practice.

Type of common-item set. Two designs for the common-item set were used in this study: MC-only (abbreviated as MC32CR0 in tables) and mixed-type (abbreviated as MC16CR3 in tables). MC-only common-item set means that the set consists of only MC items. Mixed-type common-item sets included both MC and CR items. When the FDD was .25, although the old and new forms shared 32 MC items and 3 CR items in common, in the MC-only common-item set design, only the shared 32 MC items were used as common items; however, in the mixed-type common-item set design, 16 out of 32 MC items and 3 CR items were used as common items. For the mixed-type set, the ratio of the total MC common-item score points to the total CR common-item score points was 1:1 which was consistent with the ratio between the total score points of MC section and CR section. For both types of common-item set, the total score points for the set was fixed in order to avoid having different numbers of common-item score points influence the equating results. As described earlier, the same sets of common items were also used for the condition of no form difficulty difference.

Group ability difference. In the context of item-type multidimensionality, abilities of both old form group and new form group followed bivariate normal distributions under a certain MC-CR COR (ρ) condition. In the CINEG design, the two groups might have different proficiency levels. Under each MC-CR COR condition, five levels of group ability differences, expressed as ability effect size (ES), were considered:

- ES=.00: $(\theta_{MC}, \theta_{CR})_{old} \sim BN(0,0,1,1,\rho)$ and $(\theta_{MC}, \theta_{CR})_{new} \sim BN(0,0,1,1,\rho)$

- ES=.05: $(\theta_{MC},\theta_{CR})_{old} \sim BN(0.05,0.05,1,1,\rho)$ and $(\theta_{MC},\theta_{CR})_{new} \sim BN(0,0,1,1,\rho)$
- ES=.10: $(\theta_{MC},\theta_{CR})_{old} \sim BN(0.10,0.10,1,1,\rho)$ and $(\theta_{MC},\theta_{CR})_{new} \sim BN(0,0,1,1,\rho)$
- ES=.25: $(\theta_{MC},\theta_{CR})_{old} \sim BN(0.25,0.25,1,1,\rho)$ and $(\theta_{MC},\theta_{CR})_{new} \sim BN(0,0,1,1,\rho)$
- ES=.50: $(\theta_{MC},\theta_{CR})_{old} \sim BN(0.50,0.50,1,1,\rho)$ and $(\theta_{MC},\theta_{CR})_{new} \sim BN(0,0,1,1,\rho)$

Note that, the population means for the MC and CR sections for the new form group were always set to zero and the standard deviations were all one. .00 indicated no group difference and .50 represented a large group difference. For levels of .05, .10, .25, and .50, the old form group was assumed to be more able than the new form group.

Summary of simulation conditions. Combining form difficulty difference, correlation between MC and CR sections, type of common-item set, and group ability difference, there were a total of 80 ($2\times4\times2\times5=80$) simulation conditions. These conditions are summarized in Table 1.

In addition, to indicate the extent to which the findings were stable and did not depend on the particular items chosen for the simulation, the entire simulation study was replicated, starting from the constructions of new sets of old and new test forms. Therefore, the total number of study conditions was 160 ($80 \quad 2=160$).

Data Simulation and Equating

Under each simulation condition, a sample size of 5,000 was chosen for each group so that equating would be reasonably precise (Hanson & Beguin, 2002; Kim & Lee, 2004; Kirkpatrick, 2005). Given the item parameters on the test form and ability distributions, simulated data were generated using the following steps:

(1) Randomly draw 5,000 $(\theta_{MC},\theta_{CR})$ from the given ability distribution;

(2) Generate item responses for the MC items using θ_{MC}, the given item parameters, and the 3PL model;

(3) Generate item responses for the CR items using θ_{CR}, the given CR item parameters, and the GR model;

(4) Compute composite scores and common-item set scores for examinees.

Equating was conducted to equate scores for the new form group on the score scale for the old form. The equating relationship was estimated using two methods: the presmoothed

frequency estimation method (Pre_FE) and the presmoothed chained equipercentile equating method (Pre_FE). The presmoothing procedure used with the frequency estimation method was the bivariate log-linear presmoothing. A fixed set of parameters, 6-6-1, which is commonly used in practice, was used in the bivariate log-linear presmoothing procedure. For the chained equipercentile equating method, the univariate log-linear presmoothing procedure was used. In the univariate log-linear presmoothing method, both the composite scores and common item scores were modeled using a polynomial of degree 6. When simulating data in the way that is described above, extreme score points might have zero (missing) frequencies. To deal with this situation, the equating function, beyond the score range within percentile ranks of .5 and 99.5, was estimated using linear interpolation.

In addition, in equating, the common items were used as internal common items. The new form population received a weight of one. Equating was conducted using *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009).

For each simulation condition, the simulation process was replicated 500 times. That is, for each simulation condition, 500 sets of data were generated for the old and new form groups, respectively, and then equating was performed using each pair of the data sets, which resulted in 500 estimated equating functions.

Criterion Equating Relationships

To establish a criterion equating relationship, the single group (SG) design was used. That is, the same group of examinees were assumed to take both forms. When the test forms to be equated had the same difficulty levels, the criterion equating relationship was the identity equating. When test forms being equated had different difficulties, the criterion equating function was established by using the equipercentile equating method in the SG design, based on the group taking the new form, under each MC-CR COR condition. The sample size used for establishing the criterion equating function was 1,000,000.

Evaluation Criteria

To evaluate the equating results, conditional absolute bias (*CABias*), conditional standard error of equating (*CSEE*), and conditional root mean squared error (*CRMSE*) were calculated at each raw score point. These indices were computed using the following formulas:

$$CABias(x_i) = |\bar{e}_Y(x_i) - e_Y(x_i)|\,, \qquad (1)$$

$$CSEE(x_i) = \sqrt{\frac{\sum_{r=1}^{500}\left[\hat{e}_{Y,r}(x_i) - \bar{e}_Y(x_i)\right]^2}{500 - 1}}, \quad (2)$$

$$CRMSE(x_i) = \sqrt{CABias(x_i)^2 + CSEE(x_i)^2}, \quad (3)$$

where

$$\bar{e}_Y(x_i) = \frac{\sum_{r=1}^{500} \hat{e}_{Y,r}(x_i)}{500}. \quad (4)$$

In Equations 1, 2, and 4, $\hat{e}_{Y,r}(x_i)$ is the old form equivalent score for score x_i on the new form at replication *r*; $\bar{e}_Y(x_i)$ is the mean of the old form equivalent scores for x_i over 500 replications; and $e_Y(x_i)$ is the criterion equated score for x_i.

Further, summary statistics were calculated to summarize the amount of error over the entire score scale: weighted average bias (*WABIAS*), weighted average standard error of equating (*WASEE*), and weighted average root mean squared error (*WARMSE*). The three indices were calculated using the following equations:

$$WABIAS = \sqrt{\sum_{i=0}^{K} w(x_i)[CABias(x_i)]^2}, \quad (5)$$

$$WASEE = \sqrt{\sum_{i=0}^{K} w(x_i)[CSEE(x_i)]^2}, \quad (6)$$

$$WARMSE = \sqrt{\sum_{i=0}^{K} w(x_i)[CRMSE(x_i)]^2}, \quad (7)$$

where *K* is the maximum score on the new form (Form X); $w(x_i)$ is the relative frequency of score point x_i on Form X or new form; and $w(x_i)$ is computed based on the raw score distribution for the new form population group.

To provide guidance as to the magnitude of error for acceptable equating, the difference that matters (DTM; Dorans & Feigenbaum, 1994) was used. In this study, the DTM was set to .50 score units. Following Lee, He, Hagge, Wang, and Kolen (2012), *CABias* was compared to the DTM. If *CABias* was smaller than the DTM along the score scale, the equating was considered acceptable, otherwise, it was considered unacceptable.

Results

As described in the Methods section, a simulation study and a replication of the study were completed, each of which contained 80 simulation conditions. The main purpose of conducting the replication is to assess the extent to which the findings are stable and do not depend on the particular items chosen for the simulation. In general, the patterns of the findings tended to be consistent across the two simulation studies. Therefore, only results for the first study are reported and discussed in this section.

Tables 2 to 9 provide the results for the three summary statistics (*WABIAS*, *WASEE*, and *WARMSE*) obtained using the two different common-item sets, the first four of which (Tables 2 to 5) are for the no form difficulty difference conditions and the last four (Tables 6 to 9) are for the form difficulty difference conditions. For MC32CR0, Tables 2 and 3 are for the Pre_CE method and the Pre_FE method, respectively, under the no form difficulty difference condition, whereas Tables 6 and 7 are for the two equating methods under the non-zero form difficulty difference condition. For MC16CR3, Tables 4 and 5 are for the results obtained using the two different equating methods under the no form difficulty difference condition, whereas Tables 8 and 9 are for the non-zero form difficulty difference condition.

In addition, to evaluate the amount of bias in the equating results across different score levels, absolute conditional bias was compared to the DTM for the two different common-item set designs in Figures 1 through 20. Figures 1 through 10 display the absolute conditional bias for the no form difficulty difference condition, and Figures 11 to 20 show the absolute conditional bias for the non-zero form difficulty difference condition. Figures 1 through 5 are for the Pre_CE method, one for each group ability difference. Figures 6 to 10 are for the Pre_FE method, one for each group ability difference. Among the figures for the non-zero form difficulty difference condition (Figures 11 through 20), Figures 11 through 15 are for the Pre_CE method, and the remaining five figures are for the Pre_FE method. Each figure contains four panels, one for each item-type test dimensionality. In each panel, the solid curve represents MC32CR0, and the dashed curve represents MC16CR3. In addition, the horizontal line denotes the DTM line. Figures 1 through 20 use the same figure format. Note that, in all the figures, the horizontal axis representing the new form raw composite scores was truncated at both tails where there were insufficient data to show meaningful equating results.

In this section, the results are organized based on three main variables that were manipulated: item-type test dimensionality, group ability difference, and equating method. In general, the patterns of the results are consistent across different test form difficulty levels. In addition, under which conditions the two types of common-item sets produce adequate equating results was summarized.

Item-Type Test Dimensionality

When evaluating the impact of item-type test dimensionality on the comparisons of different common-item types, the results from the two common-item sets were compared under the same form difficulty level, group ability difference level, and equating method, for each item-type test dimensionality level. Then whether consistent conclusions are obtained across different item-type test dimensionality levels is evaluated. For example, the results in Table 2 (for MC32CR0) are compared with those in Table 4 (for MC16CR3), at each dimensionality level, and then whether the conclusion varies with item-type test dimensionality was assessed. Note that the only difference between Tables 2 and 4 is the use of different common-item types. When the test is unidimensional (i.e., $\rho = 1.0$), the two common-item sets produce similar results for WABIAS, WASEE, and WARMSE. This demonstrates that, when equating unidimensional mixed-format test forms, MC-only common-item set and mixed-format common-item set perform similarly. For unidimensional test score equating, the MC-only common-item set is preferable due to the fact that there may be issues associated with the use of CR items.

When the test is multidimensional (i.e., $\rho = .90, .75, .50$), MC16CR3 produces smaller WABIAS, WASEE, and WARMSE than does MC32CR0, except for no group difference condition (ES .00). When there is no group ability difference, the two common-item sets have similar WABIAS. In addition, when the groups do not differ in ability, the two different types of common-item sets produce similar WASEE and WARMSE for tests in which the MC and CR sections are relatively highly correlated (e.g., $\rho = .90$). The findings, regarding the effect of item-type test dimensionality on the comparison of different common-item types, are consistent across different form difficulty difference levels and equating methods, such as comparing Table 3 *vs*. Table 5, Table 6 *vs*. Table 8, and Table 7 *vs*. Table 9.

In addition, as demonstrated in each table, for multidimensional tests, as the correlation between the MC section and the CR section decreases, the advantage of using the mixed-format

common-item set, MC16CR3, becomes more notable. This observation holds for different form difficulty levels and different equating methods.

Group Ability Difference

Under each group ability difference condition, the results obtained using MC32CR0 and MC16CR3 are compared at the same item-type test dimensionality level and form difficulty level, and for the same equating method. For example, compare the results in Table 2 *vs*. Table 4, Table 3 *vs*. Table 5, and so on. When the groups used in the equating do not differ in performance, $ES = .00$, the WABIAS results of the two common-item sets are very similar, which is true for each item-type test dimensionality. For WASEE, the two common-item sets result in similar results when the disattenuated correlation of MC and CR sections is moderately high to high (i.e., $\rho = 1.0, .90, .75$). When the group ability difference is relatively small but not zero (i.e., $ES = .05, .10$), the two common-item sets perform almost the same when the MC and CR sections are highly correlated (e.g., $\rho = 1.0, .90$). As the disattenuated correlation between the MC and CR sections decreases to .75 and .50, MC16CR3 performs better than MC32CR0. When there is a relatively large (e.g., .25) to large (e.g., .50) group ability difference, the mixed-type common-item set performs better than the MC-only common-item set, except for the unidimensional test. For unidimensional test score equating, the two types of common-item sets perform almost the same in terms of WABIAS.

Equating Method

Two different equating methods, PreSm_FE and PreSm_CE, were used in the present study. The findings of the effects of item-type test dimensionality and group ability difference on the comparisons of the common-item sets are consistent no matter which equating method is used.

In addition, according to the comparisons between Table 2 and Table 3, Table 4 and Table 5, etc., it can be seen that, for the same common-item set, the two equating methods produce similar WABIAS results when there is no group ability difference, regardless of item-type test dimensionality. As group ability difference increases and/or the disattenuated correlation between the MC and CR sections decreases, the difference of the WABIAS results produced by the two equating methods increases, and the PreSm_CE method performs better than the PreSm_FE method due to it having smaller WABIAS values. This finding is likely caused by the frequency estimation method being based on stronger assumptions about similarity

of examinee groups than the chained equipercentile equating method. When item-type test multidimensionality increases and/or group ability difference increases, these assumptions might be violated, and further more bias is introduced.

Summary of Conditions for Adequate Equating Results

A DTM of .50 was used as the criterion to evaluate whether adequate equating results were obtained in this study. Figures 1 through 20 display the absolute conditional bias obtained under various study conditions. When *CABias* was smaller than the DTM along the score scale, the equating was considered acceptable. Otherwise, it was considered unacceptable. The conditions under which adequate equating results were obtained are summarized as follows.

The MC-only common-item set, MC32CR0, tended to produce adequate equating results under the condition of:

1. $ES = .00$ and $ES = .05$, regardless of item-type test dimensionality and equating method;
2. $ES = .10$ and PreSm_CE, regardless of item-type test dimensionality;
3. $ES = .10$, PreSm_FE, and $\rho = 1.0 / .90$; or
4. $ES = .25$, PreSm_CE, and $\rho = 1.0$.

The mixed-type common-item set, MC16CR3, tended to produce adequate equating results under the following conditions:

1. $ES = .00$, $ES = .05$, and $ES = .10$, regardless of item-type test dimensionality or equating method; or
2. $ES = .25$ and PreSm_CE, regardless of item-type test dimensionality.

Conclusions

The selection of common items is a crucial element when equating multiple mixed-format test forms under the CINEG design, especially for equating multidimensional mixed-format tests. When different item types might contribute to the common-item set, it is critical to investigate how different types of common-item sets influence the equating results. In addition, the use of CR item(s) as common item(s) might cause some practical issues in mixed-format test score equating. Therefore, in practice, many testing programs do not use CR item(s) as common-items, which might lead to other concerns. For testing programs which use mixed-format tests, it is important to have information about the conditions in which an MC-only common-item set performs similarly as a mixed-format common-item set, and especially under what conditions an

MC-only common-item set produces equating results that are as adequate as the results for a mixed-format common-item set. This present study is an attempt to address these questions.

The results of this study show that, for both common-item sets, as group ability difference increases, bias increases, regardless of item-type test dimensionality and equating method. Item-type test dimensionality has an influence on bias only when non-zero group ability difference exists, that is, as the disattenuated correlation between MC and CR sections decreases, bias increases. For the no group ability difference condition, bias tends to be similar and small regardless of item-type test dimensionality. This finding results because, when the group ability difference is zero, the CINEG design is truly a random groups design, and in the random groups design, there is no need for a common-item set, or in another words, any common-item set can work.

Using a DTM criterion of .50, the two common-item sets produced "acceptable" equating results when no or small group ability difference exists, regardless of item-type test dimensionality. This finding was observed for both equating methods. When moderate group ability difference exists (e.g., .10), to obtain adequate equating results, the MC-only common-item set can be used with the Pre_CE method, and the disattenuated correlation between the MC and CR section can be even as low as .50.

In addition, the results of the present study demonstrate that the conclusions did not vary with different form difficulty levels. As described earlier, two simulation studies were performed, and the second one is a single replication of the entire study of the first simulation study, starting from the construction of new sets of old and new test forms. In general, the patterns of the findings tend to be consistent across the two simulation studies, which suggests that the findings were stable and did not depend on the particular items chosen for the simulation.

Several limitations of the study should be recognized. The test considered in this present study involves two different item types, MC and CR. Care should be taken when generalizing the results to tests containing more than two different item types. Another limitation is that the ability difference between groups is assumed to be the same on both MC and CR dimensions. In addition, the ability distributions of the two groups being equated are assumed to have a standard deviation of 1 on both dimensions. These two assumptions might not be characteristics of operational testing programs. Future research should examine the adequacy of equating when these assumptions are violated. This simulation uses a simple structure model in the context of

factor analysis. That is, the MC items loaded solely on one dimension whereas the CR items loaded solely on the other dimension. It would be interesting to use a more complex multidimensional IRT model in future research, such as one in which both MC and CR items load on a common factor, and where the CR items also load on a unique factor.

References

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*, 77-92.

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.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum Associates.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579-621). Westport, CT: American Council on Education/Praeger.

Hagge, S. L. (2011). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups*. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3-24.

He, Y. (2011). *Evaluating equating properties for mixed-format tests*. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Kim, S.-H., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report Series 2004-5). Iowa City: ACT.

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating*. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2). (CASMA Monograph No. 2.2)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Livingston, S. A. (2009). *Constructed-response test questions: Why we use them; how we score them*. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Manhart, J. J. (1996, April). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325-337.

Perkhounkova, Y., & Dunbar, S. B. (1999). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*, 354-379.

Reshetar, R., & Melican, G. J. (2010, April). *Design and evaluation of mixed-format large scale assessments for the Advanced Placement Program (AP)*. Paper presented at the Annual Meeting of the American Education Research Association in Denver, CO.

Sykes, C., Hou, L-L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113-123.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29–44). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H., Wang, X-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement*, *31*, 183-199.

Table 1

Simulation Conditions

Factor	Number of Levels	Description
Form Difficulty Difference	2	.00, .25
Correlation between MC and CR Sections	4	.50, .75, .90, 1.0
Type of Common-Item Set	2	MC-only (32 MC) Mixed-Type (32 MC + 3 CR)
Group Ability Difference	5	.00, .05, .10, .25, .50
Replication of the Entire Simulation Study	1	Starting from the constructions of new sets of old and new test forms
Total Number of Conditions:	$160\ (2\times4\times2\times5\times2=160)$	

Table 2

Summary Statistics for the Condition of No Form Difficulty Difference Using MC32CR0 and the Pre_CE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.013	0.063	0.119	0.305	0.605
$\rho = 0.90$	0.016	0.104	0.160	0.450	0.868
$\rho = 0.75$	0.018	0.126	0.273	0.669	1.296
$\rho = 0.50$	0.011	0.226	0.406	1.049	2.078
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.323	0.340	0.338	0.341	0.352
$\rho = 0.90$	0.380	0.380	0.376	0.386	0.408
$\rho = 0.75$	0.423	0.435	0.423	0.435	0.445
$\rho = 0.50$	0.476	0.469	0.462	0.489	0.510
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.323	0.346	0.358	0.457	0.700
$\rho = 0.90$	0.380	0.394	0.408	0.593	0.959
$\rho = 0.75$	0.423	0.453	0.504	0.798	1.370
$\rho = 0.50$	0.477	0.520	0.615	1.157	2.140

Table 3

Summary Statistics for the Condition of No Form Difficulty Difference Using MC32CR0 and the Pre_FE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.013	0.122	0.231	0.584	1.133
$\rho = 0.90$	0.012	0.180	0.324	0.853	1.637
$\rho = 0.75$	0.024	0.241	0.505	1.241	2.403
$\rho = 0.50$	0.011	0.396	0.739	1.878	3.690
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.307	0.316	0.314	0.318	0.329
$\rho = 0.90$	0.350	0.346	0.343	0.352	0.365
$\rho = 0.75$	0.380	0.389	0.382	0.384	0.393
$\rho = 0.50$	0.410	0.399	0.392	0.410	0.427
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.307	0.339	0.390	0.665	1.180
$\rho = 0.90$	0.350	0.390	0.472	0.922	1.677
$\rho = 0.75$	0.381	0.458	0.633	1.299	2.435
$\rho = 0.50$	0.410	0.563	0.836	1.922	3.714

Table 4

Summary Statistics for the Condition of No Form Difficulty Difference Using MC16CR3 and the Pre_CE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.014	0.065	0.121	0.307	0.582
$\rho = 0.90$	0.010	0.068	0.131	0.318	0.616
$\rho = 0.75$	0.015	0.075	0.146	0.337	0.660
$\rho = 0.50$	0.014	0.076	0.149	0.380	0.747
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.332	0.325	0.320	0.335	0.354
$\rho = 0.90$	0.341	0.327	0.332	0.328	0.355
$\rho = 0.75$	0.331	0.330	0.327	0.334	0.351
$\rho = 0.50$	0.323	0.324	0.327	0.333	0.356
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.332	0.331	0.342	0.454	0.681
$\rho = 0.90$	0.341	0.334	0.357	0.457	0.711
$\rho = 0.75$	0.331	0.338	0.358	0.475	0.747
$\rho = 0.50$	0.323	0.332	0.359	0.505	0.827

Table 5

Summary Statistics for the Condition of No Form Difficulty Difference Using MC16CR3 and the Pre_FE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.014	0.127	0.243	0.610	1.173
$\rho = 0.90$	0.011	0.130	0.258	0.633	1.230
$\rho = 0.75$	0.013	0.142	0.283	0.675	1.313
$\rho = 0.50$	0.014	0.155	0.300	0.759	1.480
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.314	0.306	0.305	0.315	0.332
$\rho = 0.90$	0.324	0.309	0.313	0.311	0.332
$\rho = 0.75$	0.314	0.312	0.308	0.316	0.329
$\rho = 0.50$	0.304	0.302	0.306	0.305	0.328
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.314	0.331	0.390	0.686	1.219
$\rho = 0.90$	0.324	0.336	0.405	0.705	1.274
$\rho = 0.75$	0.315	0.343	0.418	0.745	1.354
$\rho = 0.50$	0.304	0.340	0.428	0.818	1.516

Table 6

Summary Statistics for the Condition of Form Difficulty Difference Using MC32CR0 and the Pre_CE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.124	0.140	0.172	0.312	0.586
$\rho = 0.90$	0.125	0.154	0.216	0.438	0.876
$\rho = 0.75$	0.127	0.186	0.291	0.667	1.305
$\rho = 0.50$	0.117	0.245	0.444	1.069	2.118
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.327	0.327	0.326	0.336	0.353
$\rho = 0.90$	0.377	0.379	0.380	0.385	0.419
$\rho = 0.75$	0.428	0.417	0.417	0.429	0.441
$\rho = 0.50$	0.472	0.476	0.476	0.473	0.509
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.350	0.356	0.369	0.458	0.684
$\rho = 0.90$	0.397	0.409	0.437	0.583	0.971
$\rho = 0.75$	0.446	0.457	0.508	0.794	1.377
$\rho = 0.50$	0.486	0.535	0.651	1.169	2.179

Table 7

Summary Statistics for the Condition of Form Difficulty Difference Using MC32CR0 and the Pre_FE Method

Weighted Average Bias (WABIAS)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.133	0.173	0.264	0.581	1.109
$\rho = 0.90$	0.128	0.221	0.370	0.830	1.644
$\rho = 0.75$	0.128	0.281	0.515	1.233	2.414
$\rho = 0.50$	0.112	0.412	0.782	1.898	3.737
Weighted Average Standard Error of Equating (WASEE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.310	0.307	0.303	0.311	0.328
$\rho = 0.90$	0.346	0.347	0.347	0.353	0.382
$\rho = 0.75$	0.376	0.373	0.373	0.381	0.391
$\rho = 0.50$	0.414	0.406	0.407	0.415	0.428
Weighted Average Root Mean Square Error (WARMSE)					
			ES		
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.337	0.352	0.402	0.658	1.157
$\rho = 0.90$	0.369	0.411	0.507	0.902	1.688
$\rho = 0.75$	0.397	0.467	0.635	1.290	2.446
$\rho = 0.50$	0.428	0.578	0.881	1.943	3.762

Table 8

Summary Statistics for the Condition of Form Difficulty Difference Using MC16CR3 and the Pre_CE Method

Weighted Average Bias (WABIAS)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.119	0.125	0.168	0.303	0.596
$\rho = 0.90$	0.101	0.119	0.159	0.316	0.618
$\rho = 0.75$	0.101	0.110	0.175	0.346	0.673
$\rho = 0.50$	0.097	0.131	0.190	0.399	0.769
Weighted Average Standard Error of Equating (WASEE)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.328	0.336	0.322	0.331	0.344
$\rho = 0.90$	0.325	0.320	0.320	0.328	0.351
$\rho = 0.75$	0.323	0.322	0.324	0.334	0.349
$\rho = 0.50$	0.331	0.319	0.322	0.329	0.358
Weighted Average Root Mean Square Error (WARMSE)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.349	0.358	0.363	0.449	0.688
$\rho = 0.90$	0.341	0.341	0.358	0.456	0.711
$\rho = 0.75$	0.338	0.340	0.369	0.481	0.758
$\rho = 0.50$	0.345	0.345	0.374	0.517	0.848

Table 9

Summary Statistics for the Condition of Form Difficulty Difference Using MC16CR3 and the Pre_FE Method

Weighted Average Bias (WABIAS)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.124	0.163	0.265	0.594	1.177
$\rho = 0.90$	0.109	0.164	0.271	0.621	1.228
$\rho = 0.75$	0.107	0.159	0.295	0.672	1.321
$\rho = 0.50$	0.100	0.191	0.331	0.771	1.503
Weighted Average Standard Error of Equating (WASEE)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.313	0.319	0.304	0.313	0.323
$\rho = 0.90$	0.308	0.304	0.304	0.311	0.333
$\rho = 0.75$	0.303	0.302	0.303	0.312	0.326
$\rho = 0.50$	0.311	0.301	0.302	0.309	0.333
Weighted Average Root Mean Square Error (WARMSE)					
	ES				
Correlation	.00	.05	.10	.25	.50
$\rho = 1.00$	0.336	0.358	0.404	0.672	1.220
$\rho = 0.90$	0.327	0.345	0.407	0.695	1.273
$\rho = 0.75$	0.321	0.341	0.423	0.741	1.361
$\rho = 0.50$	0.326	0.356	0.448	0.830	1.540

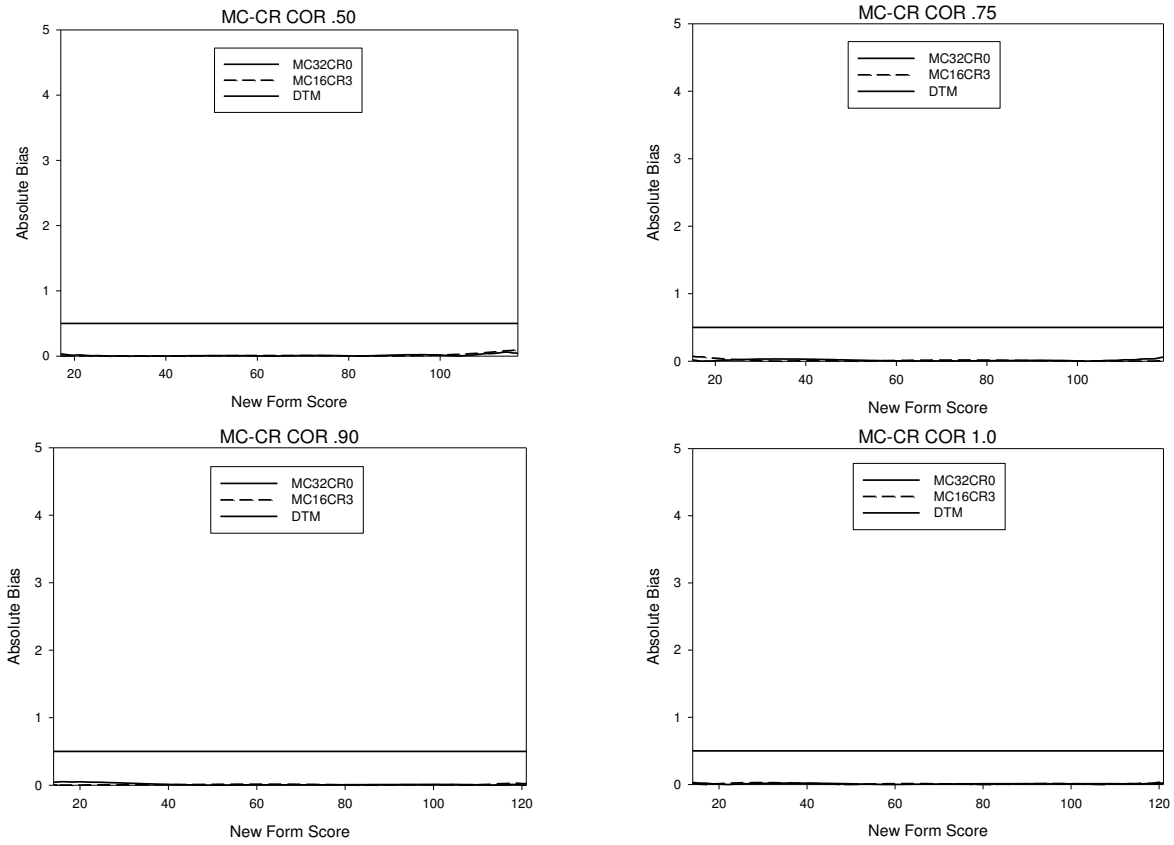


Figure 1. Absolute conditional bias for $ES = 0.00$ and Pre_CE under no form difficulty difference.

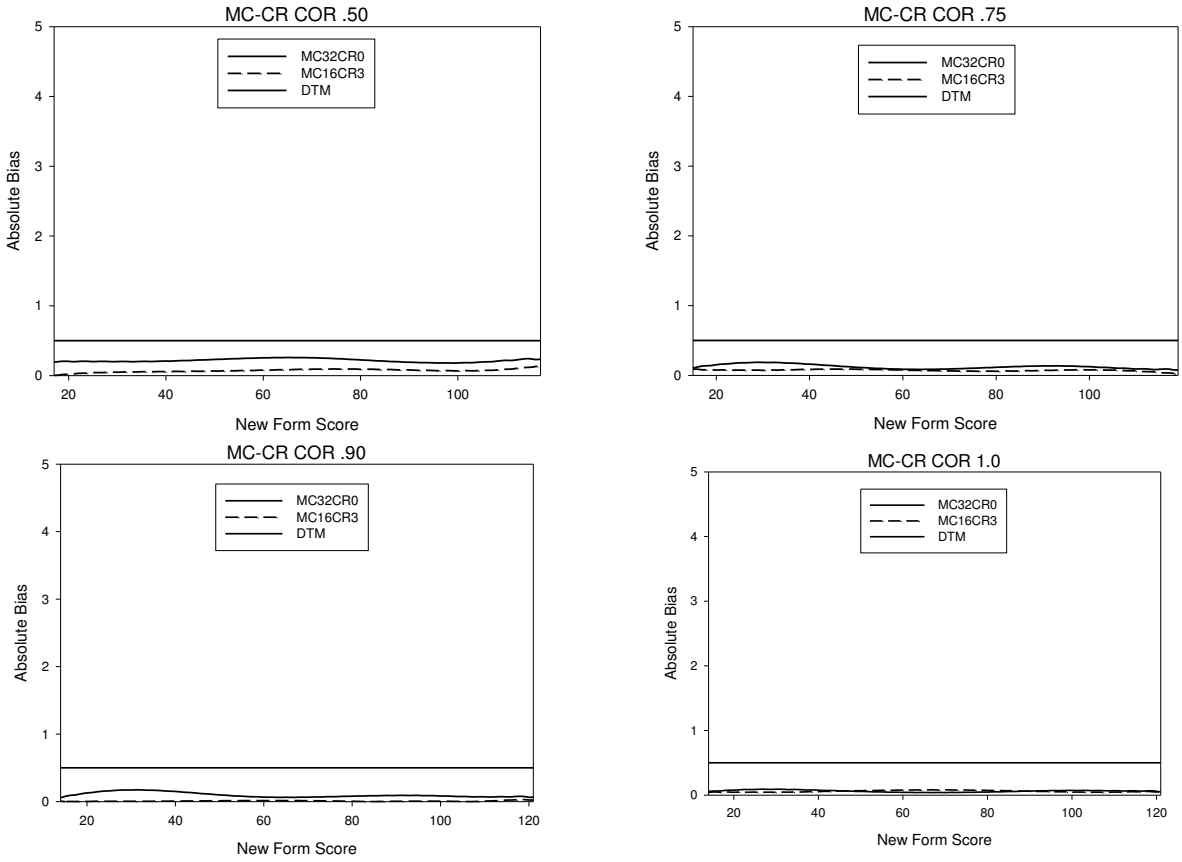


Figure 2. Absolute conditional bias for $ES = 0.05$ and Pre_CE under no form difficulty difference.

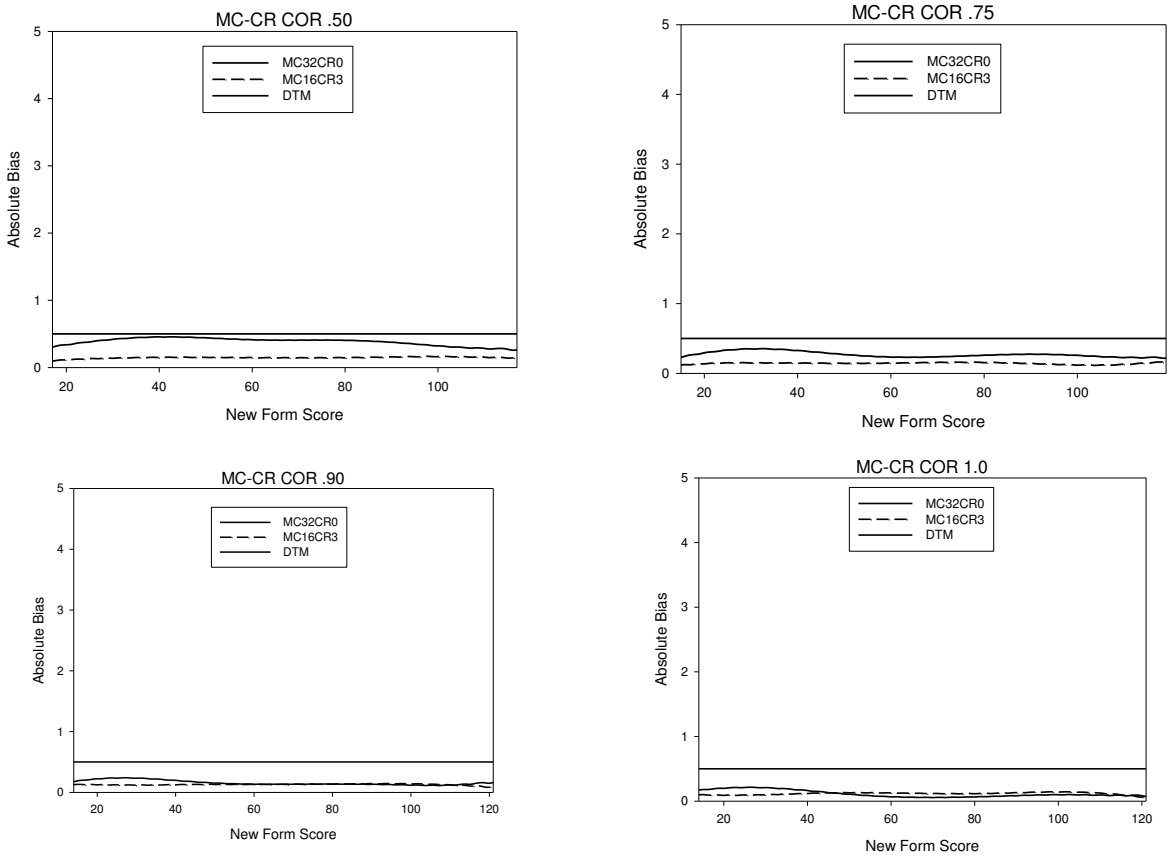


Figure 3. Absolute conditional bias for $ES = 0.10$ and Pre_CE under no form difficulty difference.

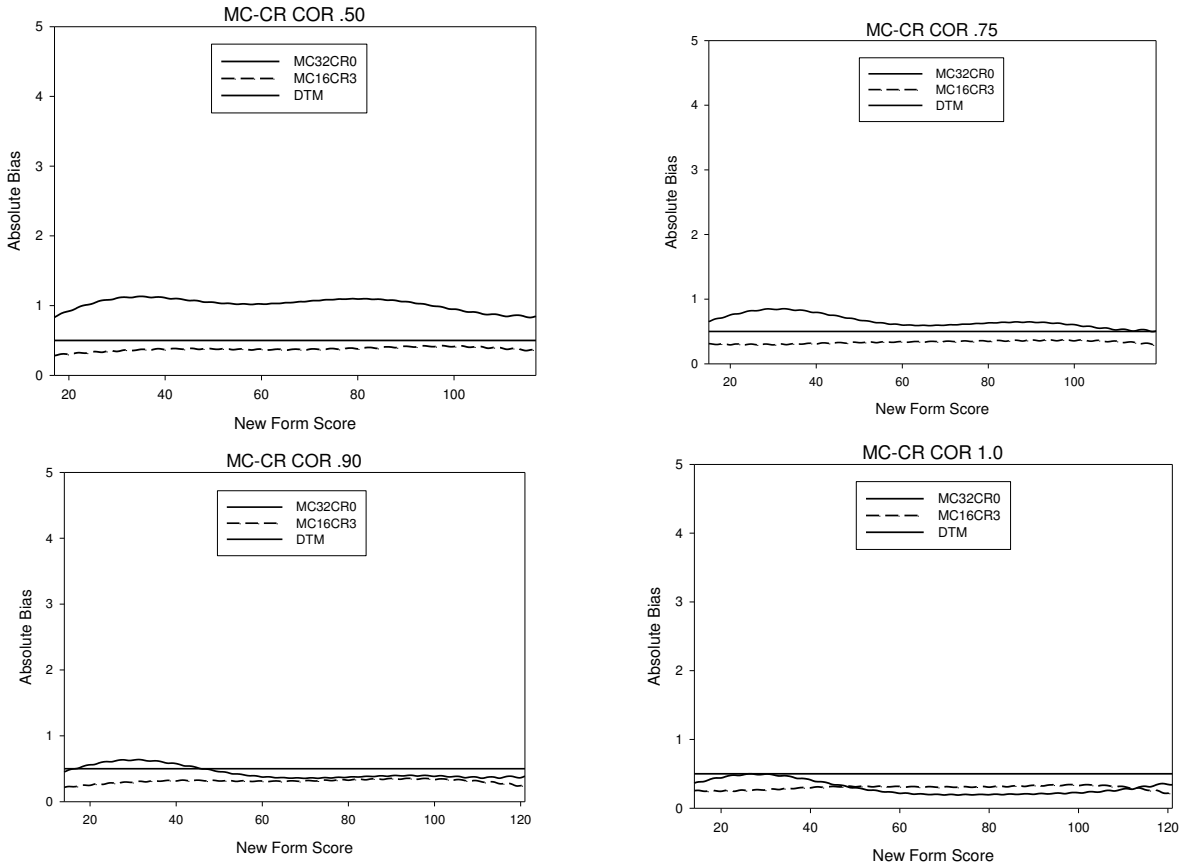


Figure 4. Absolute conditional bias for $ES = 0.25$ and Pre_CE under no form difficulty difference.

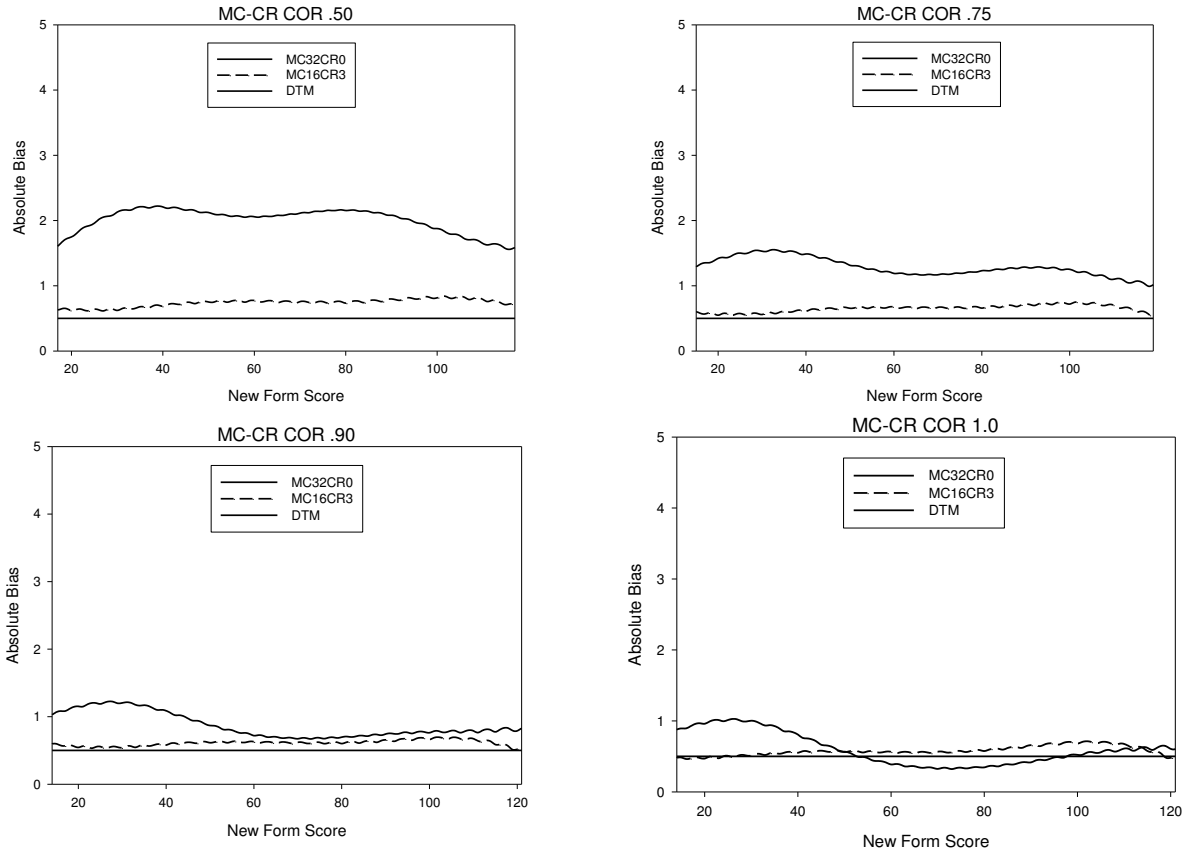


Figure 5. Absolute conditional bias for $ES = 0.50$ and Pre_CE under no form difficulty difference.

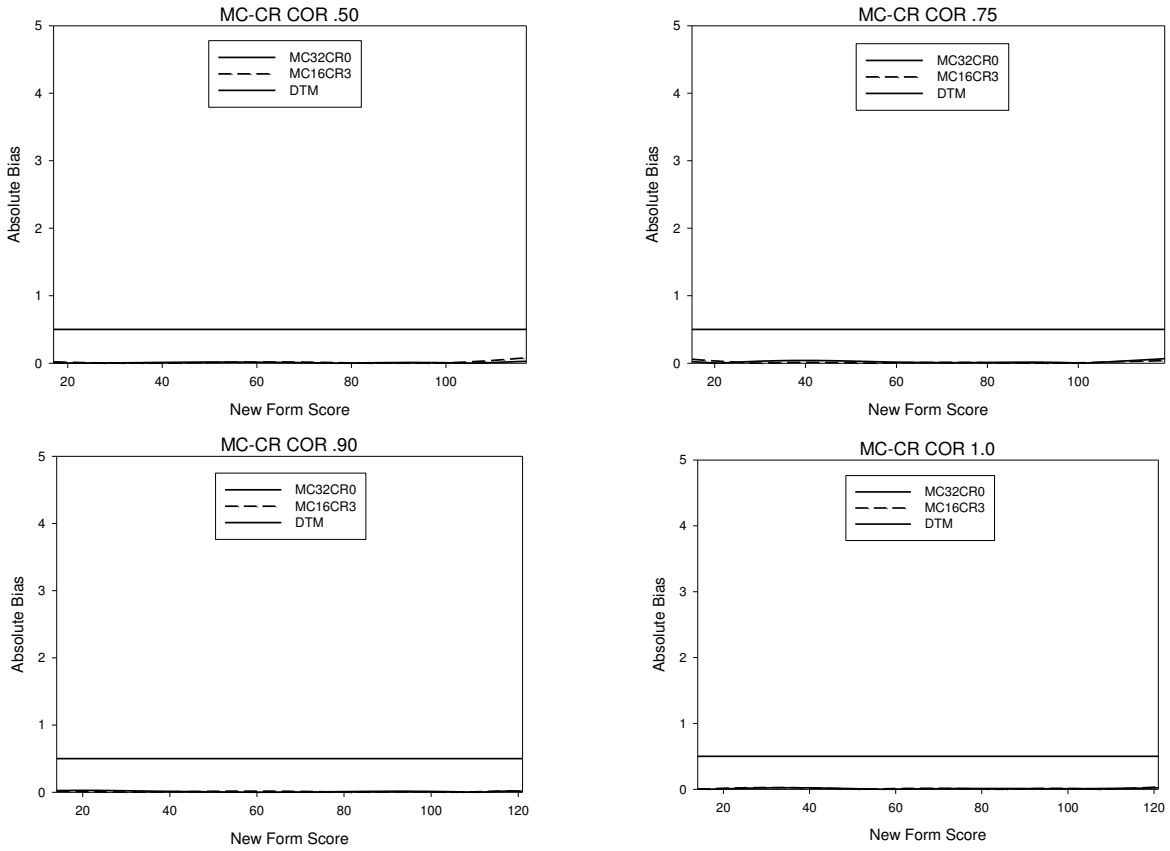


Figure 6. Absolute conditional bias for $ES = 0.00$ and Pre_FE under no form difficulty difference.

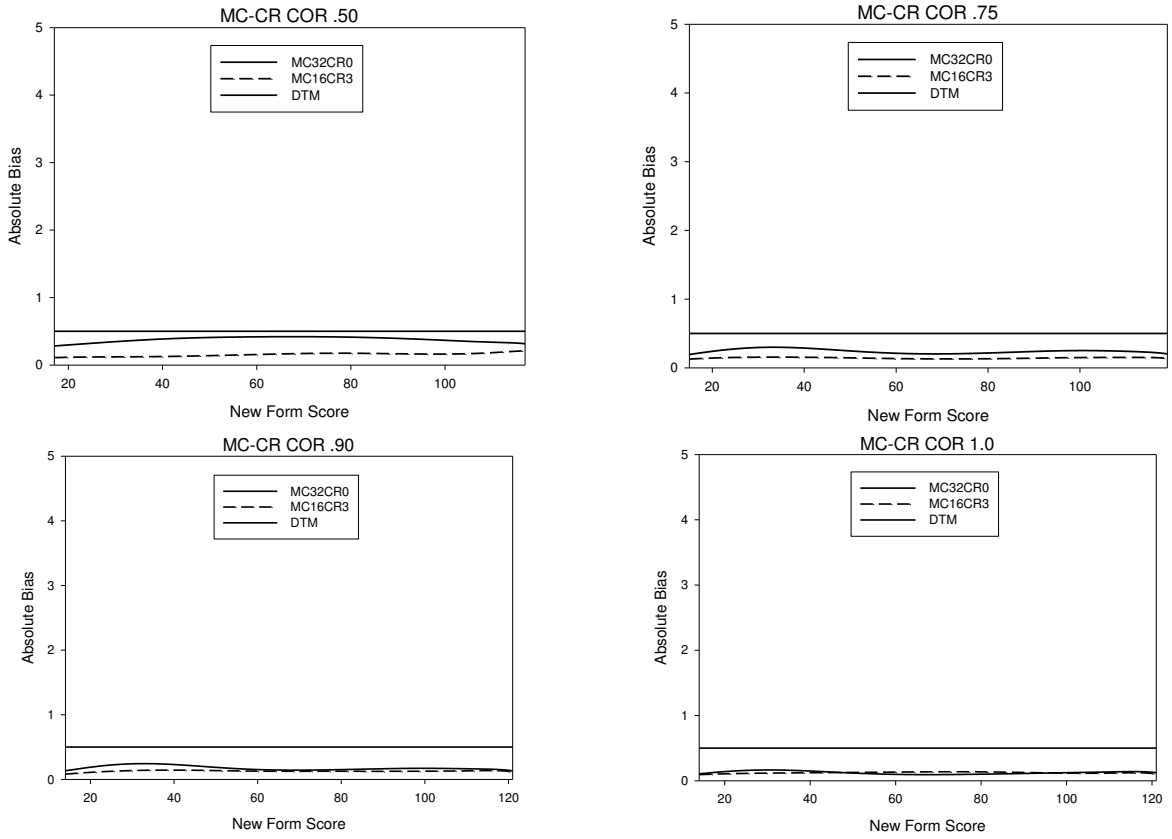


Figure 7. Absolute conditional bias for $ES = 0.05$ and Pre_FE under no form difficulty difference.

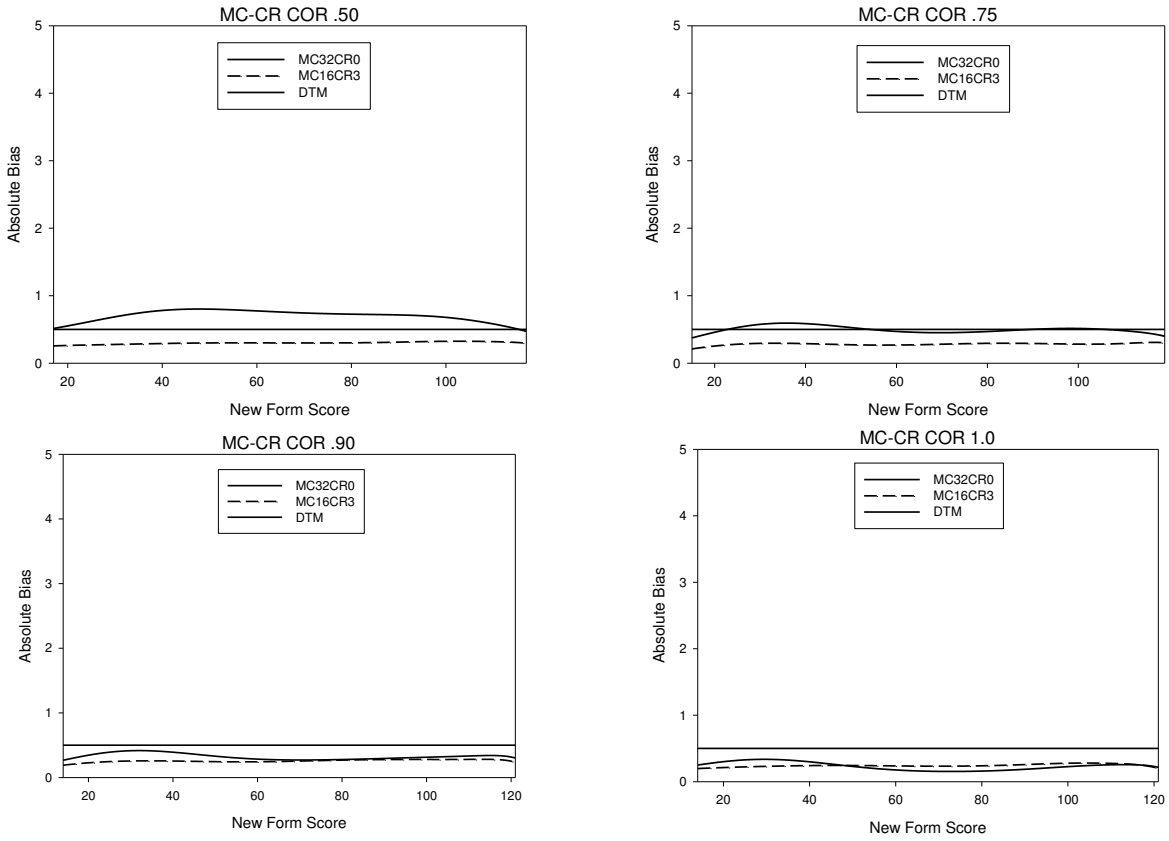


Figure 8. Absolute conditional bias for $ES = 0.10$ and Pre_FE under no form difficulty difference.

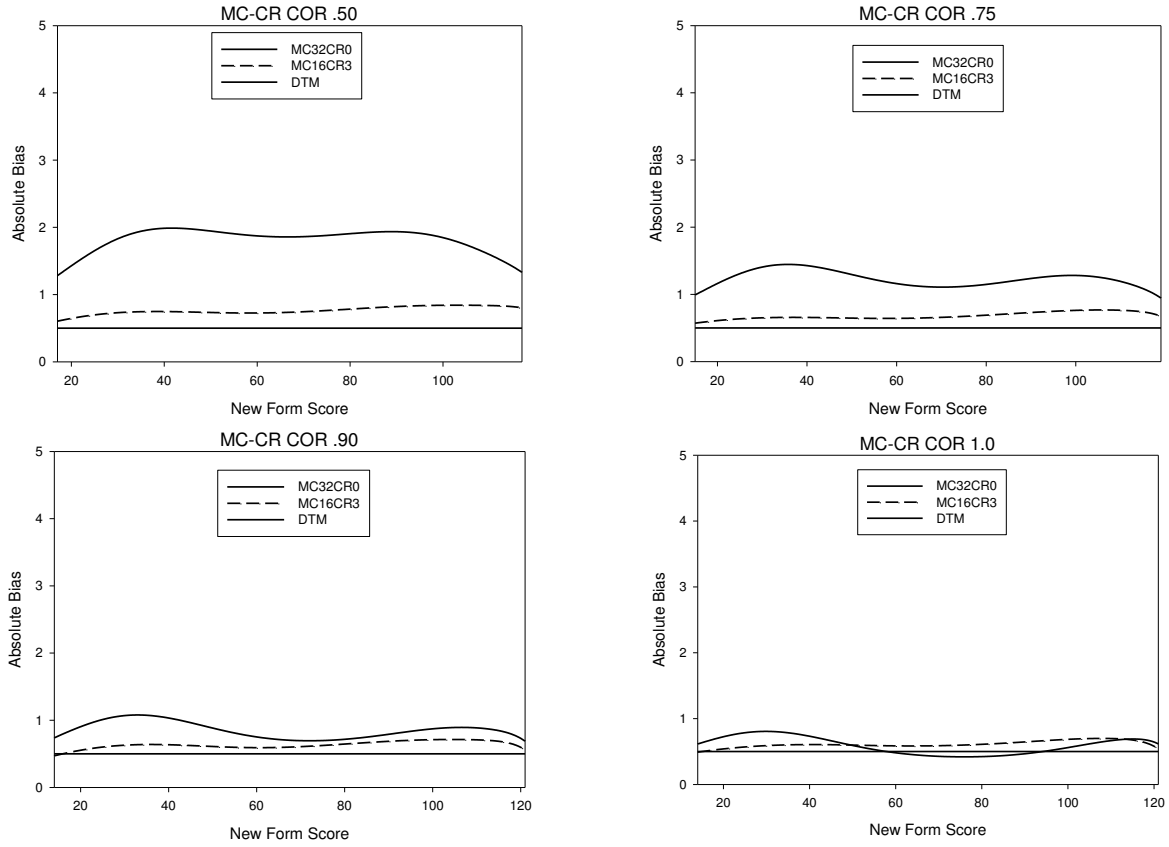


Figure 9. Absolute conditional bias for $ES = 0.25$ and Pre_FE under no form difficulty difference.

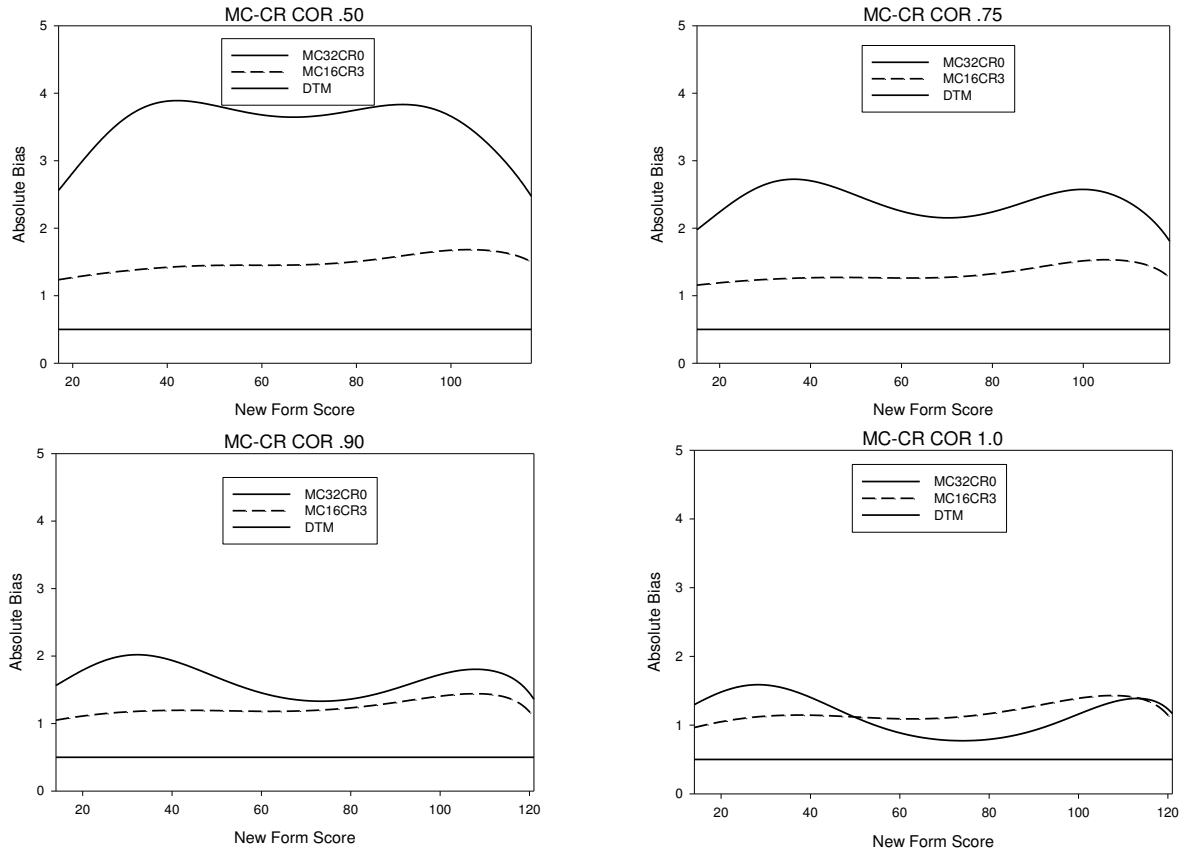


Figure 10. Absolute conditional bias for $ES = 0.50$ and Pre_FE under no form difficulty difference.

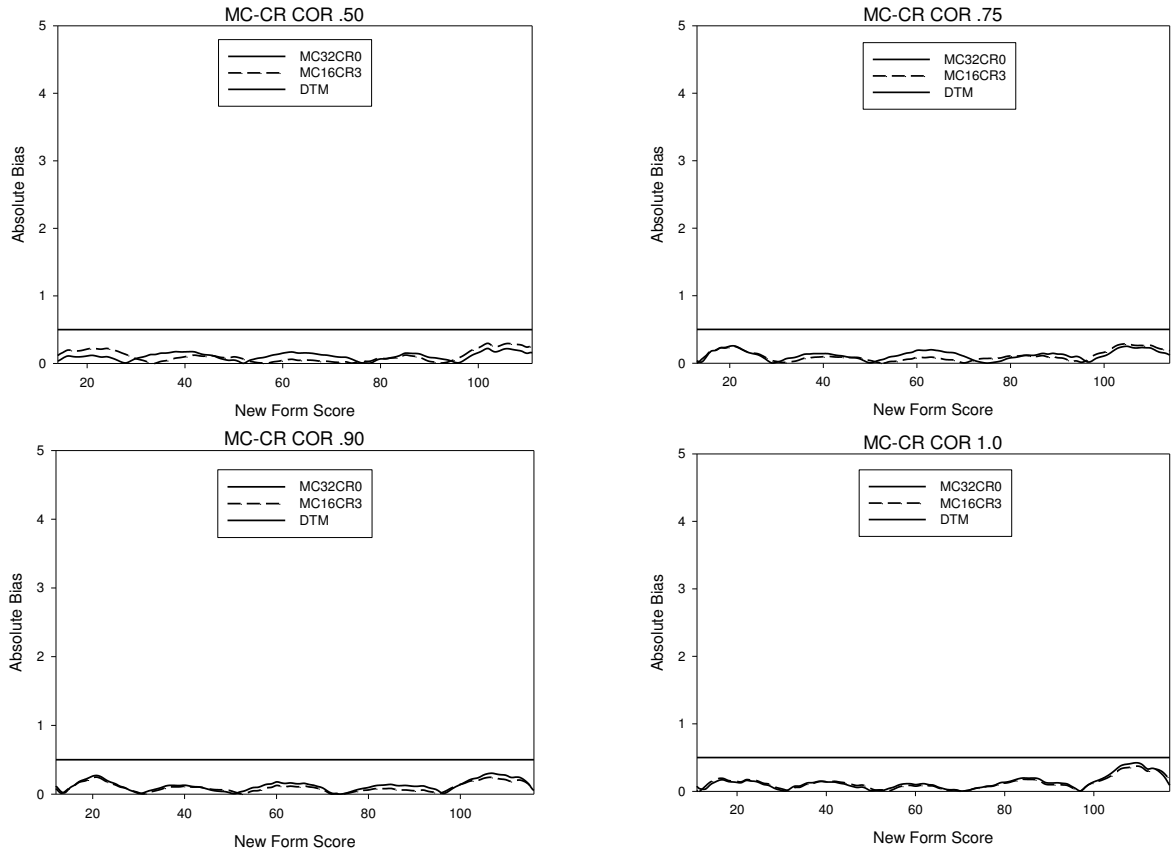


Figure 11. Absolute conditional bias for $ES = 0.00$ and Pre_CE under non-zero form difficulty difference.

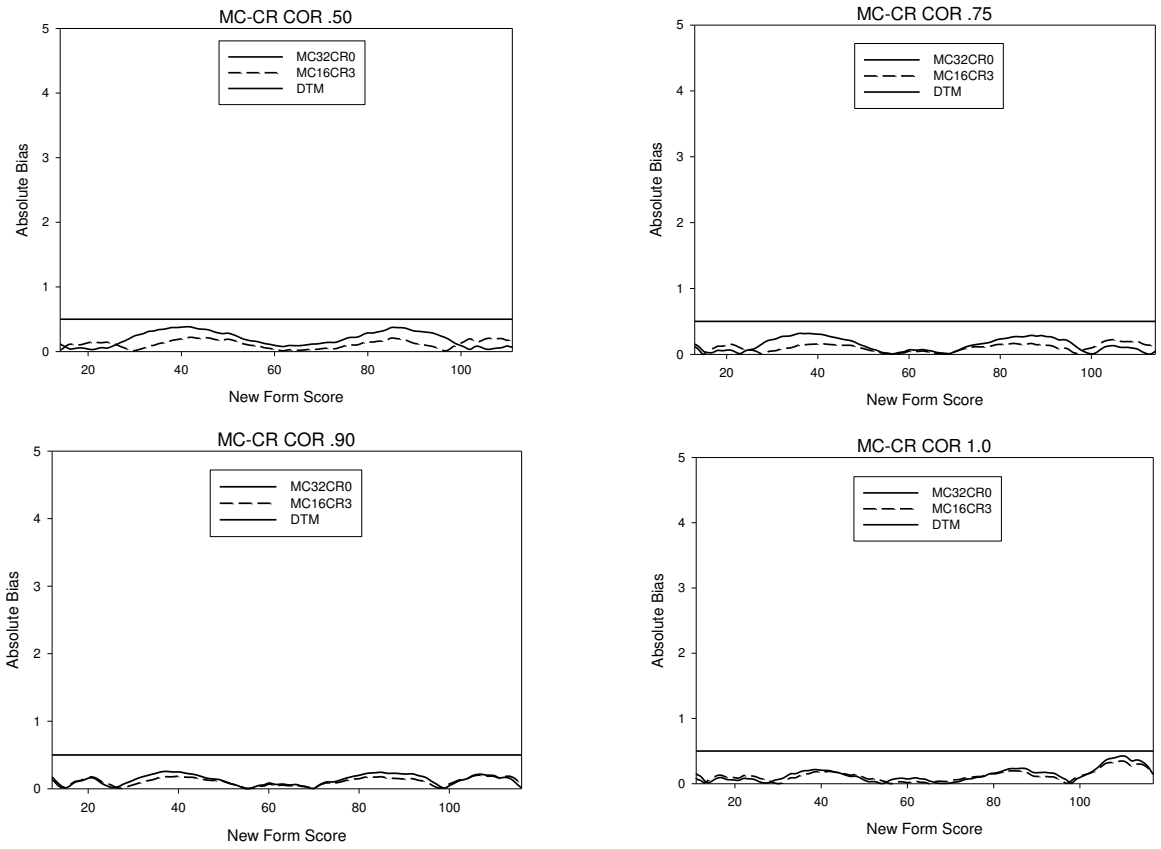


Figure 12. Absolute conditional bias for $ES = 0.05$ and Pre_CE under non-zero form difficulty difference.

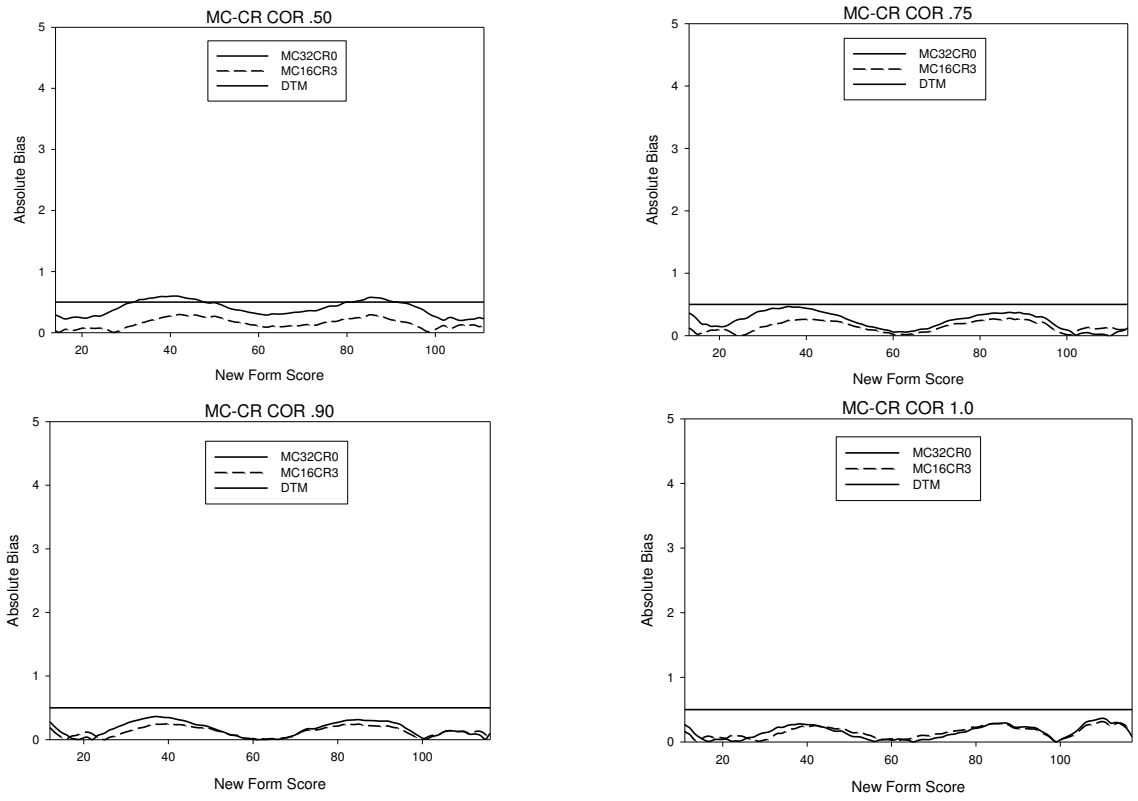


Figure 13. Absolute conditional bias for $ES = 0.10$ and Pre_CE under non-zero form difficulty difference.

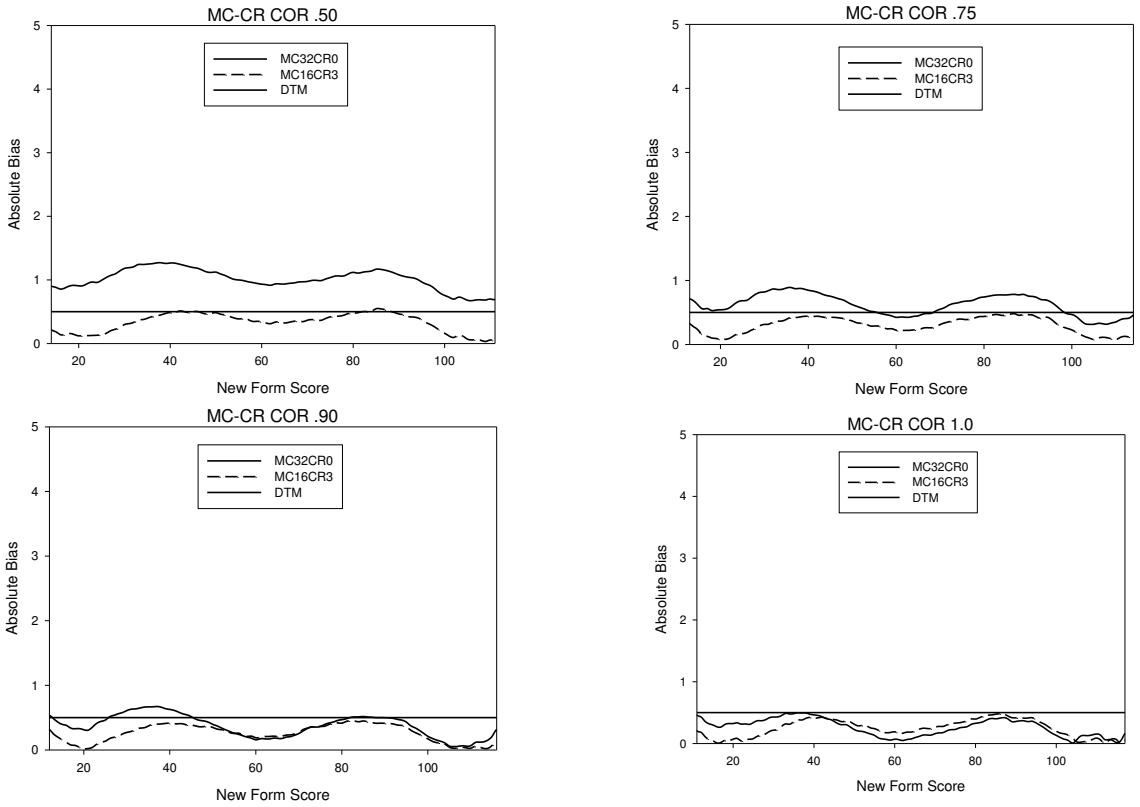


Figure 14. Absolute conditional bias for $ES = 0.25$ and Pre_CE under non-zero form difficulty difference.

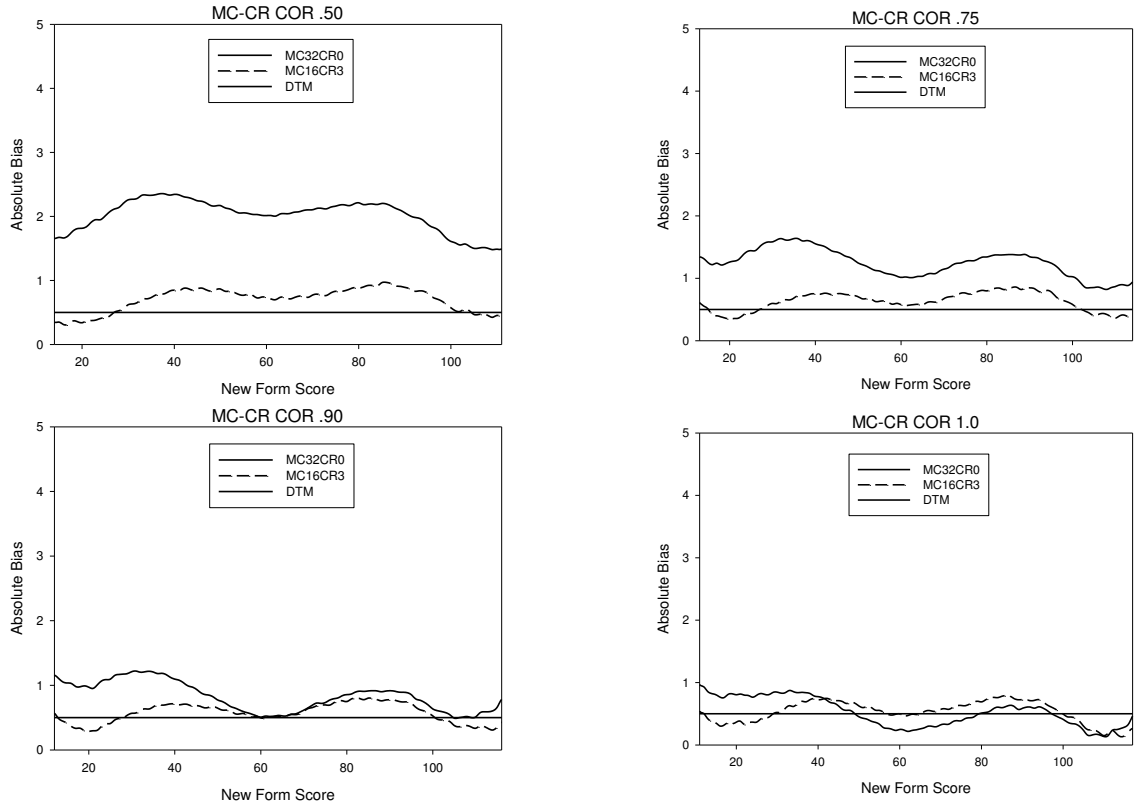


Figure 15. Absolute conditional bias for $ES = 0.50$ and Pre_CE under non-zero form difficulty difference.

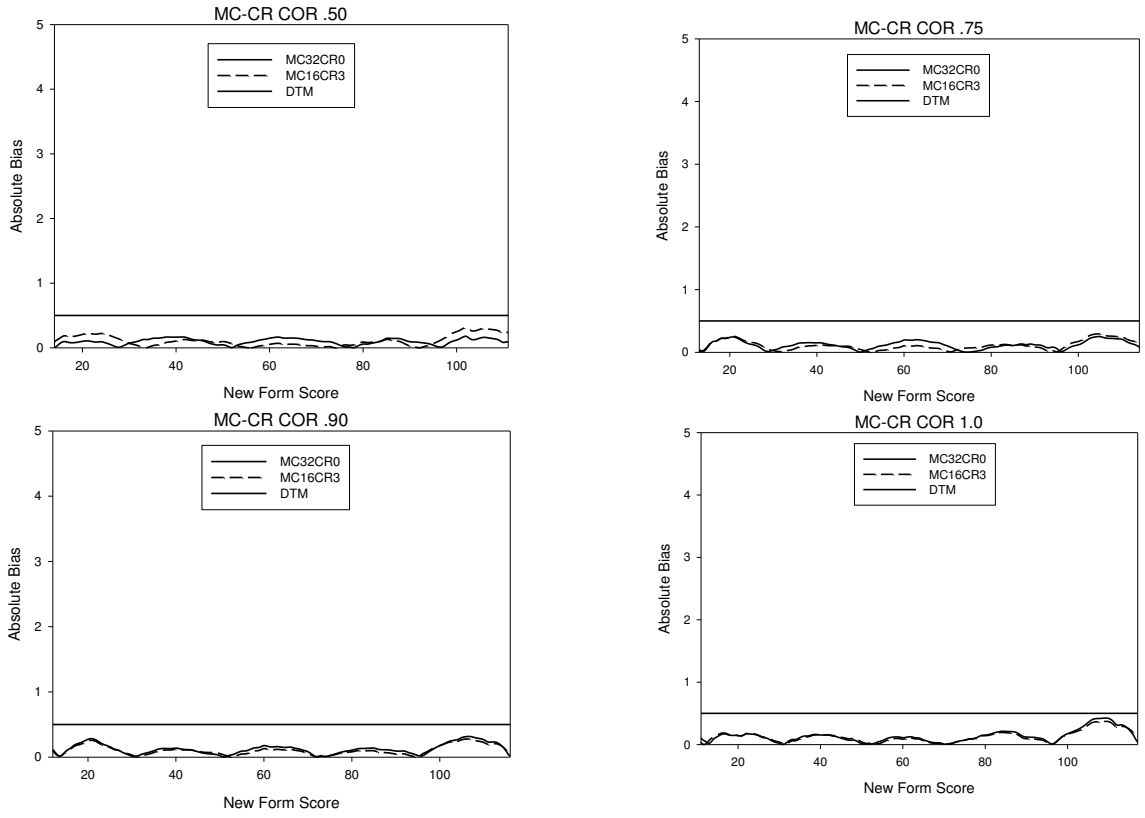


Figure 16. Absolute conditional bias for $ES = 0.00$ and Pre_FE under non-zero form difficulty difference.

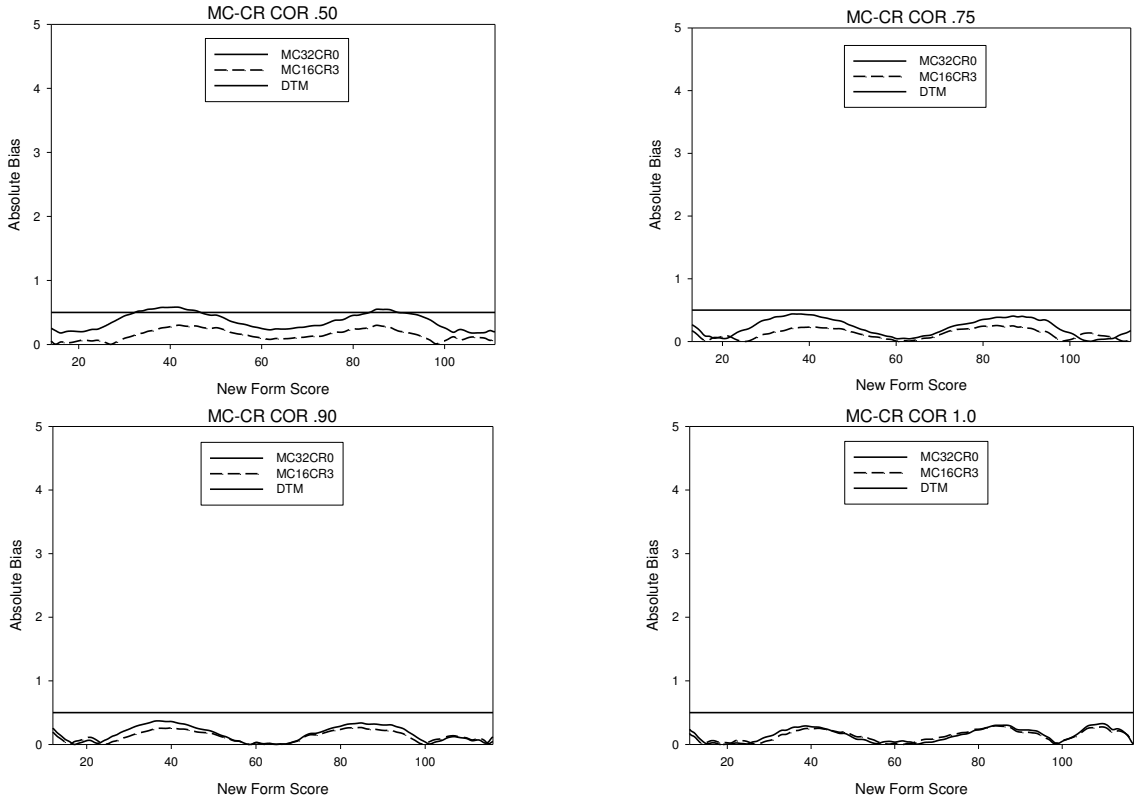


Figure 17. Absolute conditional bias for $ES = 0.05$ and Pre_FE under non-zero form difficulty difference.

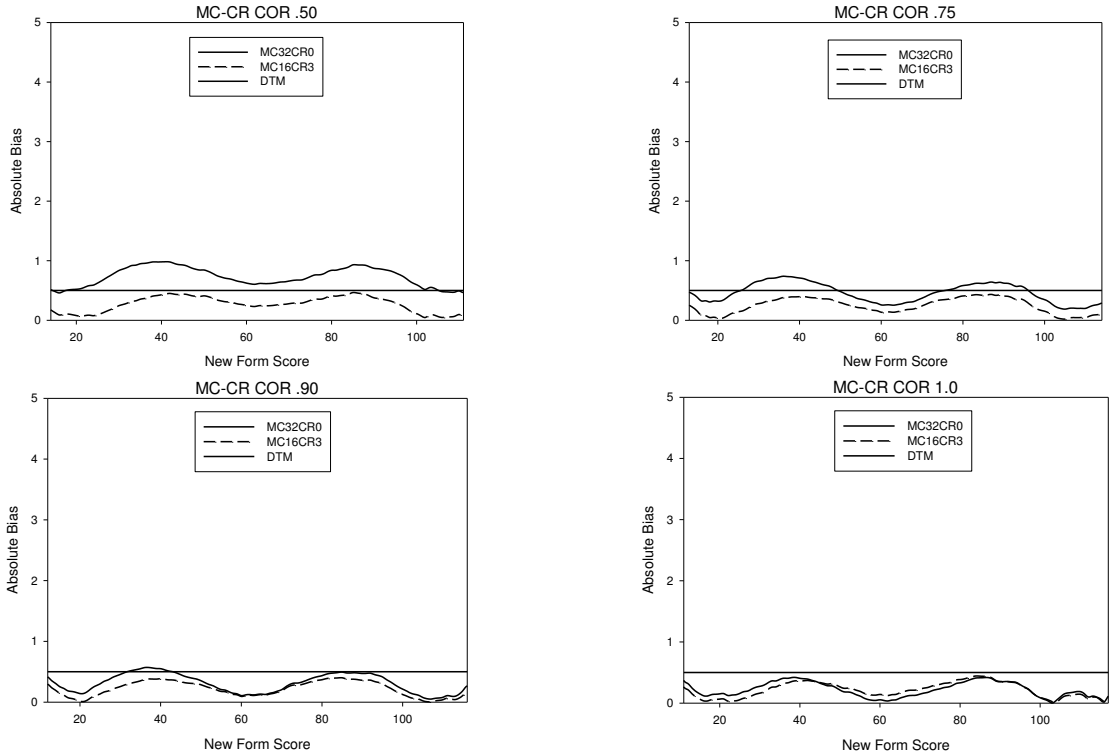


Figure 18. Absolute conditional bias for $ES = 0.10$ and Pre_FE under non-zero form difficulty difference.

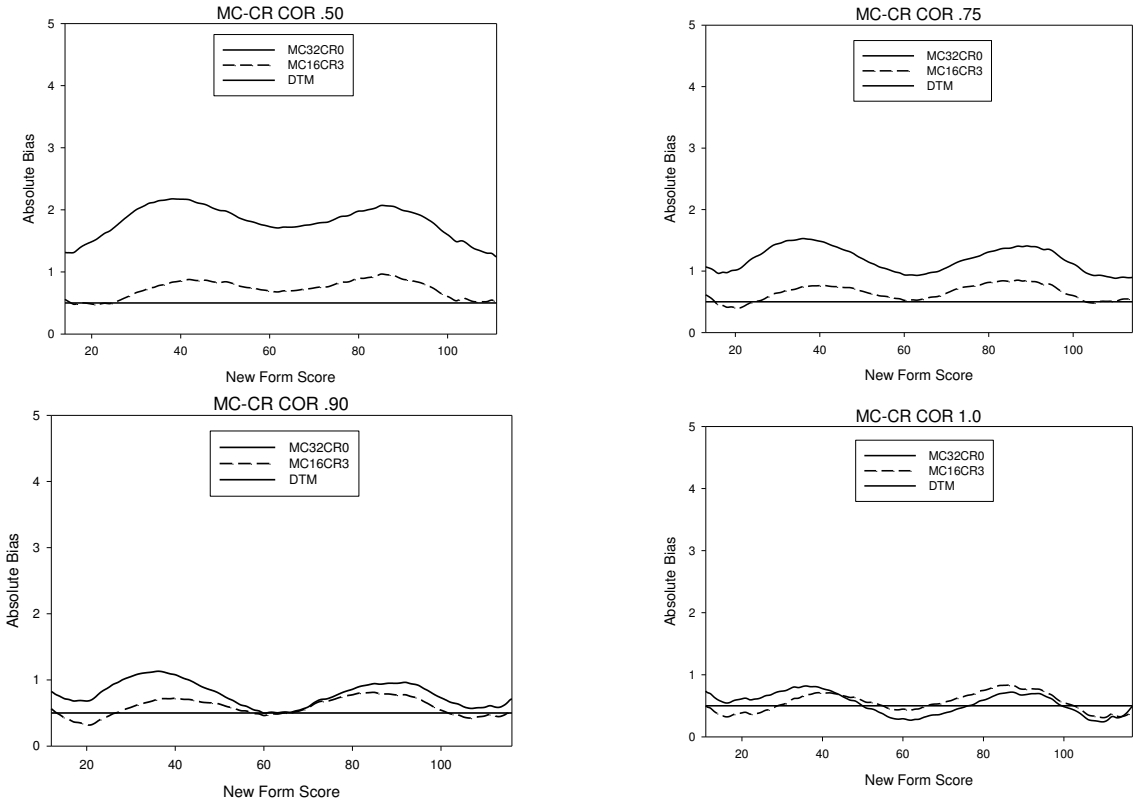


Figure 19. Absolute conditional bias for $ES = 0.25$ and Pre_FE under non-zero form difficulty difference.

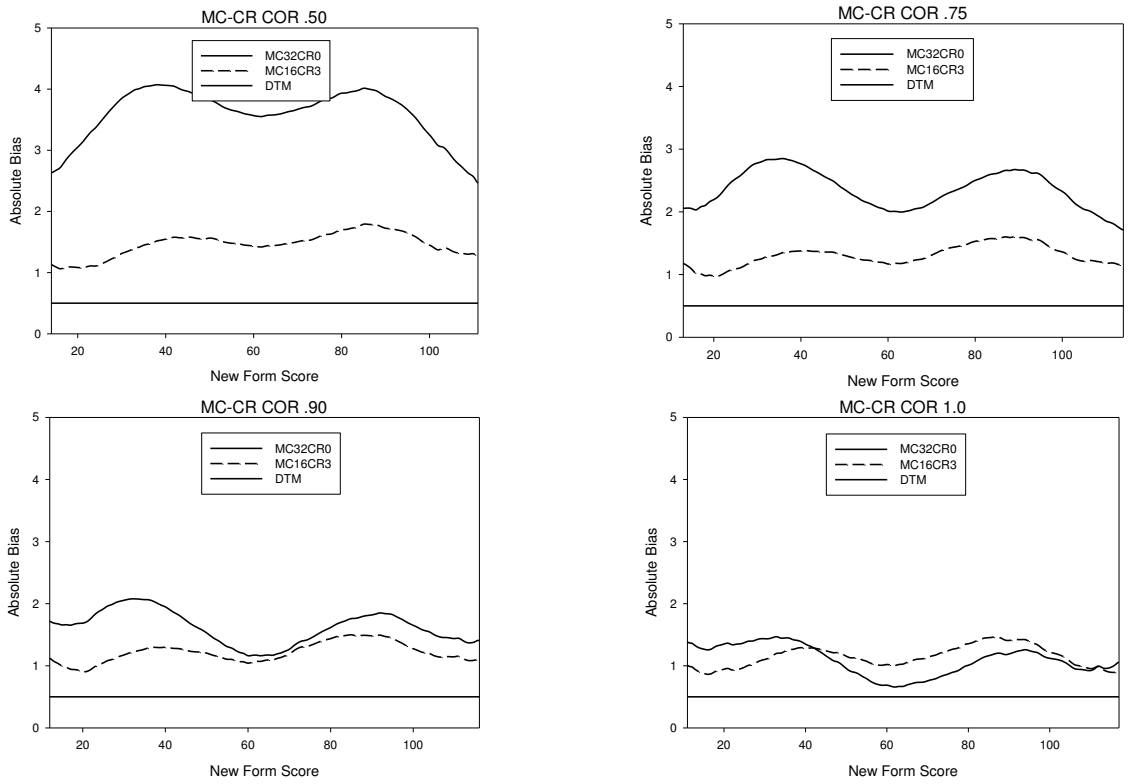


Figure 20. Absolute conditional bias for $ES = 0.50$ and Pre_FE under non-zero form difficulty difference.

Chapter 4: A Comparison of Several Item Response Theory Software Programs for Calibrating Mixed-Format Exams

Jaime Peterson, Mengyao Zhang, Seohong Pak, Shichao Wang,
Wei Wang, Won-Chan Lee, and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

This study compared differences in item parameter estimates, TCCs and TIFs that resulted from using different software programs. Specifically, two newer programs along with two well-established programs were included in the study: MULTILOG, PARSCALE, IRTPRO, and flexMIRT. In addition to making comparisons between programs, comparisons were also made within each program using three different item prior settings. In general, a trend existed such that when the 2PL model was used in comparison to the 3PL model, fewer differences were found between programs and within programs. A high degree of similarity was found between MULTILOG and IRTPRO/flexMIRT for all conditions, whereas results from PARSCALE were sometimes quite different.

A Comparison of Several Item Response Theory Software Programs for Calibrating Mixed-Format Exams

In practice, there is no direct way to measure the "correctness" of any given item response theory (IRT) model since the "truth" cannot be known. Often, fit indices are used to test the assumption of good model-data fit, with the expectation that one model provides superior fit over an alternative model. Similarly, the accuracy of different IRT calibration programs is difficult to determine when used with real data. When choosing between IRT calibration programs, decisions are generally made based on practical considerations such as flexibility, cost, and prior experience. Usually, differences observed in item and person estimates between calibration programs are not that large, but they do exist. Furthermore, differences are likely to be more pronounced for models that contain more parameters, such as when the three parameter logistic model is used in comparison to the two parameter logistic model. Similarly, the administration of mixed-format tests requires that polytomous IRT models be used for free response (FR) items, which generally contain more parameters in comparison to dichotomous models used with multiple-choice (MC) items. Since IRT model and IRT software selections have the potential to affect examinee outcomes, it is important to understand the differences associated with these decisions. Therefore, the purpose of this study was to examine differences in estimated item and test characteristics of mixed-format tests resulting from the use of different IRT models and different IRT calibration programs.

More specifically, the main objective in this study was addressed by evaluating differences in item parameter estimates, test characteristic curves (TCCs), and test information functions (TIFs) resulting from the use of different IRT software programs. This issue has become more central in the past few years with the introduction of two software programs, flexMIRT (Cai, 2012) and IRTPRO (Cai, Thissen, & du Toit, 2011). The current study compared results provided by these two newer software programs with results provided by two well-established programs that have been researched more extensively – MULTILOG (Thissen, Chen, & Bock, 2003) and PARSCALE (Muraki & Bock, 2003). This software comparison was intended to provide information to testing programs that can help inform decisions of whether adopting new software might be beneficial given the resources required.

Specifically, four different IRT model combinations were investigated. The MC items were estimated using the two- and three-parameter logistic models, and the FR items were

estimated using Samejima's (1969) Graded Response (GR) model and Muraki's (1992) Generalized Partial Credit (GPC) model. These comparisons were made using data from the AP Biology main form from 2011, which is a mixed-format exam and consists of both multiple-choice (MC) and free-response (FR) items. In addition, the effect of using the default estimation setting as well as several different item prior settings is also considered within each calibration program. Comparisons are made with respect to item parameter estimates, test characteristic curves (TCCs), and test information functions (TIFs).

Background Information

The data used in the current study was from a mixed-format test and therefore separate IRT models were required for the MC and FR items. For the MC items, the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) model were examined. For the FR items, Samejima's (1969) graded response (GR) model and Muraki's (1992) generalized partial credit (GPC) model were used. One exception was with MULTILOG, in which Bock's Nominal (BN) model was specified and contrast matrices were used to transform results into the GPC model. It should be noted that the GPC model is a special case of the BN model, so this transformation was not expected to affect results to a large extent. In total, four IRT model combinations were used: 2PL + GR, 3PL + GR, 2PL + GPC/BN, and 3PL + GPC/BN.[1]

Computer Software Parameterizations and Transformations

This section provides additional information on specific parameterizations used by each of the four software programs during item calibration. Information is provided only when the software parameterization differs from traditional parameterizations of the 2PL, 3PL, GR, and GPC/BN models, as described in Kolen and Brennan (2014). With some programs, such as flexMIRT, transformations were used to go from program-specific to traditional parameterizations, in order to make more straightforward comparisons. In such cases where transformations were required, the subsequent equations are provided in this section.

PARSCALE. In PARSCALE, the GR model is parameterized as

$$P_k(\theta) = \frac{1}{1+\exp[-Da(\theta - b + c_k)]} - \frac{1}{1+\exp[-Da(\theta - b + c_{k+1})]},$$

[1] The GPC model was used with PARSCALE, IRTPRO, and flexMIRT. The BN model was used with MULTILOG.

where *a* represents the slope; *b* is the item location; and c_k is the category boundary parameter of the k^{th} category. In order to obtain the category boundary parameters consistent with the GR model provided by Kolen and Brennan (2014), the transformation $b = b\text{-}c_k$ was used.

MULTILOG. The software program, MULTILOG does not directly provide the Muraki parameterization for the GPC model, but instead provides output in the form of Bock's (1972) nominal model. However, the GPC model can be viewed as a special case of Bock's nominal model and therefore was used in this study. Specifically, Bock's nominal model can be expressed as

$$P_{jk}(\theta) = \frac{\exp[Da_{jk}(\theta - b_{jk})]}{\sum_{u=1}^{m_j} \exp[Da_{ju}(\theta - b_{ju})]},$$

where *D* is a scaling constant; m_j is the number of response categories of item j; a_{jk} and b_{jk} are the discrimination and difficulty parameters, respectively, associated with the kth category of item j ($k = 1, \ldots, m_j$); and u is used to sum over the m_j categories in the denominator. Bock's nominal model becomes the GPC model when a_{jk} in the previous equation is replaced by $T_{jk}a_j$:

$$P_{jk}(\theta) = \frac{\exp[DT_{jk}a_j(\theta - b_{jk})]}{\sum_{u=1}^{m_j} \exp[DT_{ju}a_j(\theta - b_{ju})]},$$

with the additional requirement that T_j must be a linear vector (see Childs & Chen, 1999, for more details). In MULTILOG, the GPC model can be expressed as

$$P_{jk}(\theta) = \frac{\exp(Da_{jk}\theta - c_{jk})}{\sum_{u=1}^{m_j} \exp(Da_{ju}\theta - c_{ju})}.$$

In this parameterization, $c_{jk} = -a_{jk}b_{jk}$, and T_j is handled by contrasts. In effect, using MULTILOG to estimate GPC model item parameters demands the estimation of contrasts among the parameters: TMATRIX and FIX commands must be included, and these contrasts must also be specified for the c_{jk} parameters. Specifically, to get the GPC model parameterization, the T-matrices of triangle contrasts is used (Toit, 2003, for more details). In triangle T-matrices, the constraint $\sum c = 0$ is replaced with the constraint $c_1 = 0$, when used for the vector c. For the GPC model, MULTILOG forces $D = 1$, so the a parameters should be divided by 1.7 to be in the logistic metric. For other parameters, following computations are needed (identical procedure used for flexMIRT parameterization):

$$b = -\frac{gamma1}{slope} = -\frac{c_m}{slope*(m-1)},$$

$$\text{Fix } d_1 = 0 \text{ and compute } d_k = \frac{c_k - c_{k-1}}{slope} - \frac{c_m}{slope*(m-1)},\ k = 2, \ldots, m.$$

Additionally, different scaling constants were used in MULTILOG: by default, $D = 1.7$ for the 3PL model whereas $D = 1$ for the 2PL, GPC, and GR models.

IRTPRO. Version 2.1 of IRTPRO was used in the current study and was designed to be efficient for the estimation of item and person parameters under the multidimensional IRT framework. As a result, the parameterization used in the program is in the slope-intercept form, $a\theta + c$, as well as the traditional form, $a(\theta - b)$. The trace line for the 2PL model is expressed as,

$$T = \frac{1}{1+\exp[-(a\theta+c)]} = \frac{1}{1+\exp[-a(\theta-b)]},$$

where a is the item discrimination parameter; b is the item location parameter; c is the intercept parameter; and θ is the person latent ability parameter. The middle term of the equation is referred to as the slope-intercept parameterization and is used during the parameter estimation process. The models are always in the logistic ($D = 1$) metric. Discrimination parameter estimates can be made comparable to the normal ogive metric by dividing the IRTPRO discrimination estimates by 1.7, which was done in the current study.

flexMIRT. Similar to IRTPRO, flexMIRT (Version 1.88) was designed to handle both unidimensional and multidimensional IRT frameworks, but only provides output in the slope/intercept format for certain models. It should be stressed that a new version of flexMIRT has since been published and may resolve some of these issues. To obtain the parameters commonly used and referenced in Kolen and Brennan (2014), some parameter transformations were needed and are described for each model.

3PL model. To calculate the traditional 3PL item discrimination (a), difficulty (b), and pseudo-guessing (c) parameters, the following transformations were required:

$$a = \frac{slope}{1.7},$$

$$b = -\frac{intercept}{slope}, \text{and } c = \frac{1}{1+\exp(-logit\ guessing)} = \frac{\exp(logit\ guessing)}{\exp(logit\ guessing)+1}.$$

GR model. In the parameter output file, category intercepts (c_{k}) and an item-level slope are provided for the GR model. In order to obtain item discrimination (a) and category difficulty (b_{k}) parameters, the following transformations are required:

$$a = \frac{slope}{1.7}, \text{and}$$

$$b_{\mathrm{k}} = -\frac{c_k}{slope}, k = 1, 2, \ldots, m-1, \text{ where } m \text{ equals the number of categories.}$$

GPC model. flexMIRT uses a parameterization for the GPC model which differs dramatically from the commonly used parameterization involving item discrimination (a), item difficulty (b), and category difficulty (d_{k}). In the parameter output file, an item-level slope and category gamma parameters are given. The following transformations are necessary (refer to Thissen, Cai, & Bock, 2010, for more details):

- Suppose there are m categories. Define a $m \times (m-1)$ matrix as

$$\mathbf{T_F} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & f_{22} & \cdots & f_{2(\mathrm{m}-1)} \\ 2 & f_{32} & \cdots & f_{3(\mathrm{m}-1)} \\ \vdots & \vdots & & \vdots \\ m-1 & 0 & \cdots & 0 \end{bmatrix},$$

where $f_{\mathrm{ki}} = \sin[\pi(i-1)(k-1)/(m-1)]$, for $k = 1, \ldots, m$ and $i = 1, \ldots, m-1$.

Set $c_1 = 0$ and compute $\begin{bmatrix} c_2 \\ \vdots \\ c_m \end{bmatrix} = \mathbf{T_F} \begin{bmatrix} gamma1 \\ \vdots \\ gamma(m-1) \end{bmatrix}$.

$$a = \frac{slope}{1.7},$$

$$b = -\frac{gamma1}{slope} = -\frac{c_m}{slope*(m-1)}, \text{ and}$$

Fix $d_1 = 0$ and compute $d_{\mathrm{k}} = \frac{c_{\mathrm{k}} - c_{\mathrm{k}-1}}{slope} - \frac{c_m}{slope*(m-1)}$, $k = 2, \ldots, m$.

Method

Data

The 2011 AP Biology main form, which is a mixed-format exam, was used in the current study. Items were calibrated using four IRT model combinations and four IRT software programs, resulting in a total of 16 combinations in the study design. The main form consisted of 100 MC items with 5 options each, and 4 FR items. The MC items were dichotomously scored as 0 or 1, and FR items were scored by human raters on a 0-10 scale. The main form was administered operationally to 179,506 examinees. However, due to sample size restrictions of MULTILOG, a random sample of 3,000 examinees was generated in SAS, and used to conduct the following analyses.

Originally, analyses were conducted using the 2011 AP Biology main and alternate forms. However, the findings were generally the same for both forms, and therefore only results for the main form are presented in order to avoid redundancy.

Item Calibration

Four IRT software programs were used to conduct item calibrations: MULTILOG, PARSCALE, IRTPRO, and flexMIRT. Within each program, three item prior settings were used to estimate each of the four IRT model combinations and are referred to as the 1) baseline (with a- and c-priors), 2) c-prior only, and 3) a- and c-prior setting (Table 1). These particular item prior settings were chosen because they were compatible or could be closely approximated across all calibration programs, and are found frequently in the literature (Hanson & Beguin, 1999; Rupp, 2003). However, the selected prior settings were not fully compatible with PARSCALE and therefore an approximation was used. In PARSCALE, only the beta distribution is available for the *c*-parameter prior, whereas for MULTILOG, only a logit-normal distribution is available. For IRTPRO and flexMIRT, both beta and logit-normal distributions are available options for the c-parameter prior. As a result, several beta distributions were plotted against the normal distribution used in IRTPRO and MULTILOG, and the closest approximation was chosen (*beta* (5, 20)) and used in PARSCALE.

For all conditions, the calibration settings were intended to be as similar as possible across the four IRT software programs and can be seen in Table 2. With PARSCALE, an additional command, "GPARM = (.2)" was required in order to set the initial values for the c-parameter to 0.2 rather than using the default value of 0. When "GPARM" was omitted, results from PARSCALE differed considerably more from the IRTPRO/flexMIRT results in comparison to when it was included. This is because the initial value of the c-parameter in IRTPRO/flexMIRT is also 0.2 by default.

Characteristic Curves and Information Functions

Test characteristic curves (TCCs) and test information functions (TIFs) were computed for all conditions. The test information functions were found by summing over the 104 item information functions at 49 evenly spaced quadrature points between thetas of -6 and 6. The steps used to compute the category response functions and item information functions for the FR items can be found in the Appendix.

Results

Results were compared in terms of the estimated item parameters, characteristic curves, and information functions. It should be noted that the GPC estimates from MULTILOG were obtained by specifying the BN model with several matrix contrasts, as outlined by Childs and

Chen (1999). Across all conditions, the results from IRTPRO and flexMIRT were indistinguishable from one another. Therefore, an arbitrary decision was made to present results for only IRTPRO in an effort to simplify comparisons. It is important to stress that the decision to present results for IRTPRO and not flexMIRT was completely arbitrary and does not suggest one program is superior to the other.

MC Item Parameter Estimates

In this section, MC item parameter estimates are first compared between IRT programs and then within each program using different calibration settings. Using each model combination, between-program comparisons are made using scatter plots and results from the baseline calibration setting only. The designation of a baseline setting was made to facilitate comparisons and was arbitrary. Therefore, any of the three item prior settings could have been used as the baseline condition. For between-program comparisons, IRTPRO served as the benchmark program for comparative purposes. However, this choice was also arbitrary and does not indicate that IRTPRO should be regarded as superior to any other program.

For within-program comparisons, results from the baseline setting are plotted against those from the two alternate calibration settings, and are provided for all IRT model combinations using each program. The two alternate calibration settings differ from the baseline setting with respect to the item prior distributions only. Details concerning these differences can be found in Table 2 and apply to all IRT programs.

Comparison between IRT calibration programs. Scatter plots were used to provide a visual comparison of the differences among item parameter estimates resulting from the use of different IRT software programs. In the scatter plots, the horizontal axis represents item estimates from IRTPRO, and the vertical axis represents estimates from MULTILOG and PARSCALE. Results from the baseline setting were used to make between-program comparisons. As mentioned earlier, flexMIRT estimates were not included because they were identical to those from IRTPRO. Furthermore, scatter plots of FR items were not included because trends were similar to those found for the MC items and were also difficult to display in a succinct manner.

The differences in MC slope estimates between IRTPRO, MULTILOG, and PARSCALE for all model combinations and using the baseline settings can be found in Figure 1. The biggest difference between MULTILOG and PARSCALE slope estimates in comparison to IRTPRO

slope estimates was found with the 3PL model combinations. For all model combination, slope estimates from PARSCALE were most different from IRTPRO estimates and tended to be larger, whereas there was a high degree of consistency between the MULTILOG and IRTPRO slope estimates. The location estimates from IRTPRO, MULTILOG, and PARSCALE for all model combinations were very similar and can be found in Figure 2. Under the 3PL + GR model combination, there was one item for which the estimated location was much lower in PARSCALE in comparison to the other programs. In general, the pseudo-guessing estimates were very similar across programs, and were particularly similar between IRTPRO and MULTILOG. A visual comparison of the pseudo-guessing estimates across programs can be found in Figure 3. The greatest difference between IRTPRO and PARSCALE was found among the items with smaller c-parameters, in general. For these items, the c-parameter estimates from PARSCALE were consistently smaller than those from IRTPRO. However, differences were generally small, with the exception of one item under the 3PL + GR model combination. For this item, the difference was likely related to convergence issues with PARSCALE, as the c-parameter estimate from PARSCALE and IRTPRO were 0.0 and 0.10, respectively.

Comparison within IRT calibration programs. Differences in item parameter estimates resulting from the use of three different item prior distribution settings were compared across model combinations, and within each software program. It should be noted that within-program comparisons are not presented for flexMIRT because of the similarities found between it and IRTPRO.

For IRTPRO, the different item prior settings, resulted in virtually the same slope and location estimates for the 2PL + GPC and 2PL + GR model combinations (Figures 4 and 5). When the 3PL model was used, differences resulting from the use of different item priors were found in the slope (Figure 4) and pseudo-guessing estimates (Figure 6), but not location estimates. However, differences were more pronounced for the slope estimates than the pseudo-guessing estimates. For the slope estimates, the same trends were seen for the 3PL + GR and 3PL + GPC models – using a prior only on the c-parameter resulted in estimates that were more different from the baseline estimates in comparison to when priors were used on the a- and c-parameters. This makes sense seeing that the baseline estimates were obtained by specifying priors on the a- and c-parameters.

Differences in the slope, location, and pseudo-guessing parameter estimates resulting from the use of different prior settings in MULTILOG can be found in Figures 7 to 9, respectively. In general, the same trend was seen in the MULTILOG estimates as was found with the IRTPRO estimates. The use of different priors had very little effect on the parameter estimates when the 2PL model was used for the MC items. When the 3PL model was used with the MC items, the slope estimates varied the most as a result of using different item prior settings, the pseudo-guessing estimates varied slightly, and the location estimates were generally consistent.

A different pattern of results were found in the item parameter estimates from PARSCALE (in comparison to IRTPRO/MULTILOG) as a result of using different prior settings. Differences in the slope, location, and pseudo-guessing estimates between the baseline and other item prior settings can be found in Figures 10 to 12, respectively. Similar to the findings for IRTPRO/MULTILOG, when the 2PL model was used in PARSCALE, the slope and location estimates were very similar. However, when the 3PL model was used, the effect of using different priors looked different for PARSCALE than it did for the other programs. In general, the slope, location, and pseudo-guessing estimates remained consistent even though slightly different prior settings were used. However, there were a few items (approximately 5), for which differences were found in the location and pseudo-guessing estimates as a result of using different prior settings. For these items, it appeared as if the estimates from the baseline setting were most different from the estimates from the other two settings.

Characteristic Curves and Information Functions

In this section, TCCs and TIFs are computed for each model combination, within each software program, and using three different item prior settings. Plots of TIFs and TCCs are used to assess whether differences in item parameter estimates carried over into meaningful differences at the test level. Results are presented first for between-program comparisons using the baseline setting and then for within-program comparisons using the three prior distribution settings. Between-program comparisons are made with reference to results from IRTPRO, and within-program comparisons are made in reference to the baseline calibration setting.

Differences between IRT software programs. In general, the TCCs of each model combinations were very similar for all software programs and can be seen in Figure 13. Comparisons were made using the baseline calibration setting. Across model combinations, the

TCCs from MULTILOG and IRTPRO were nearly identical. The TCCs from PARSCALE were very similar to MULTILOG and IRTPRO, however very small differences were seen and were most noticeable for the 3PL + GR model combination.

Unlike with the TCCs, the TIFs from PARSCALE were considerably more different from the TIFs from MULTILOG and IRTPRO (Figure 14). For all model combinations the TIFs from PARSCALE had the highest peaks, suggesting greater information. The TIFs from MULTILOG and IRTPRO were basically identical to one another. For each model combination, the modes of the TIFs from each program were generally in the same location and the overall shapes of the TIFs were consistent across programs.

Differences within IRT software programs. For each software program, estimated TCCs pertaining to each model combination were plotted using results from the three prior distribution settings. The TCCs for IRTPRO, MULTILOG, and PARSCALE can be found in Figures 15 to 17, respectively. The same pattern was found for each software program – the TCCs were nearly identical even when different prior setting were used. This trend was consistent across all model combinations and software programs.

The effect of using different prior settings on the TIFs was generally different for each program, however some patterns were consistent across programs. For IRTPRO and MULTILOG the TIFs were nearly identical across the three different item prior settings when the 2PL model was used (Figures 18 and 19, respectively). For PARSCALE, the peak of the TIFs for the 2PL combinations was the highest with the baseline prior setting, followed next by the other setting that used priors on the a- and c-parameters, and was the lowest for the c-prior only setting (Figure 20).

When the 3PL model was used with the MC items, the patterns in TIFs varied across software programs, but more consistencies were seen between MULTILOG and IRTPRO. In general, for these programs, the c-prior only setting resulted in the highest TIF peak for both 3PL model combinations, whereas the baseline setting resulted in the lowest peak, and the a- and c-prior setting fell in between the two. One exception in this trend was found with the 3PL + GPC model combination for IRTPRO where the c-prior only setting still resulted in the highest maximum information, but the other two settings resulted in the same TIFs (Figure 18). For PARSCALE, the use of different prior settings on the TIFs for the 3PL model combinations followed the same trend as with the 2PL model combinations (Figure 20). More specifically, the

mode of the TIFs were the highest for the baseline setting, followed by the a- and c-prior setting, and next by the c-prior only setting. However, the TIFs corresponding to the a- and c-prior setting and c-prior only setting were much more similar (almost indistinguishable) for the 3PL model combinations, whereas they were more distinct for the 2PL combinations.

Discussion

The goal of the current study was to compare differences in item parameter estimates, TCCs and TIFs that resulted from using different software programs. Specifically, two newer programs along with two well-established programs were included in the study: MULTILOG, PARSCALE, IRTPRO, and flexMIRT. When looking at differences across programs, comparisons were typically made in reference to results produced from IRTPRO since it is a newer program. It should be noted that an arbitrary decision was made to use IRTPRO as the reference program and that any of the four programs could have been selected. Another decision was made to exclude results from flexMIRT after it was discovered that they were identical to those from IRTPRO. The consistency between these programs was not surprising since they have overlapping theoretical frameworks and authors. Therefore, in this study, results from IRTPRO were generalized to flexMIRT without compromising accuracy. However, it is important to note that calibration settings for IRTPRO and flexMIRT were made identical in this study, but the default settings of each program actually vary slightly. Therefore, if the default settings were used in each program, there is no reason to assume that results would be identical.

In addition to making comparisons between programs, comparisons were also made within each program using three different item prior settings. It was important to research the effect of different item priors because priors are often used to resolve convergence problems. Within each program, the effect of using different prior settings was investigated for each model combination and evaluated based on item parameter estimates, TCCs, and TIFs. In order to make comparisons, one of the three item prior settings was arbitrarily chosen to serve as the reference condition and was referred to as the "baseline" setting. This decision was arbitrary and was made for convenience purposes only.

This study was originally designed with the inclusion of a fourth calibration setting, which corresponded to the default setting of each program. Therefore, each model combination would have been estimated using four calibration settings within each software program. The default setting may be used frequently as a first attempt to calibrate data and therefore it is

important to understand how similar results are across software programs when it is used. However, when calibrations were conducting using default settings there were several conditions that did not converge. More specifically, none of the model combinations converged in PARSCALE, and the 3PL model combinations did not converge in IRTPRO/flexMIRT. Due to the high incidence of non-convergence, a decision was made to drop all default setting conditions from the current study.

Comparisons Between IRT Software Programs

The similarities between MULTILOG, PARSCALE, and IRTPRO took on different patterns depending on whether item parameter estimates, TCCs, or TIFs were being considered. In general, a trend existed such that when the 2PL model was used in comparison to the 3PL model, fewer differences were found between programs. For model combinations involving the 2PL models, the item estimates were very consistent across programs. However, when the 3PL model was used, greater differences emerged, especially between PARSCALE and the other programs. Comparisons for all (a, b, and c) item parameter estimates showed a high degree of similarity between IRTPRO and MULTILOG when the 3PL model was used. When PARSCALE was used, the a- and c-parameter estimates tended to differ from IRTPRO, with differences in the a-parameter being more apparent. The similarity between IRTPRO and MULTILOG may be related to the fact that for the 3PL model, both programs use the logistic metric by default. In order to make results comparable to the normal ogive metric, the a-parameter needs to be divided by a scaling constant (i.e., $D = 1.7$). Therefore, in IRTPRO and MULTILOG the prior distributions on the a-parameters are on the logistic metric, whereas in PARSCALE they are on the normal metric.

The pattern seen in the item parameter estimates did not carry over into the TCCs and TIFs. For example, the TCCs from all programs were basically identical for each model combination. When the TIFs were examined, a somewhat different pattern emerged among the IRT programs than what was found with item parameter estimates or TCCs. For each model combination, the mode of the TIF was always higher for PARSCALE, whereas the TIFs from MULTILOG and IRTPRO were basically identical.

Comparisons Within IRT Software Programs

Similar to the between-program comparisons, greater differences were found within each program as a result of using different prior settings when the 3PL model was included. For each

program, changing the item prior distributions did not affect the item parameter estimates for the 2PL model combinations.

For the 3PL model combinations, the same trends in item parameter estimates were found for IRTPRO and MULTILOG. More specifically, the use of different priors affected the a- and c-parameter estimates, but not the b-parameter estimates. For these two programs, the biggest difference was found with the a-parameter estimates such that the estimates were more similar between the two conditions that used both a- and c-priors versus the condition that used a prior distribution on the c-parameter only. Therefore, it seems that parameter estimates will be more similar in situations where priors are used on the same item parameters, even if the distributions for those priors differ, than if one of the prior distributions is removed (i.e., the c-prior only condition). In general, the use of different prior distributions had less of an impact on the 3PL parameter estimates from PARSCALE, in comparison to the other programs. Again, the reason for this may be related to the difference in scales between IRTPRO/MULTILOG and PARSCALE when the 3PL model is used.

Surprisingly, there was virtually no effect on the TCCs from using different prior settings. This trend was observed for each program and across all model combinations. When the 2PL model was used, the use of different prior settings had no effect on the TIFs from IRTPRO and MULTILOG. For these two programs, but for the 3PL model combinations, the highest TIF mode corresponded to the setting using a c-prior only, followed next by the a- and c-prior setting, and the lowest mode was found for the baseline setting. The reverse pattern was found with all model combinations in PARSCALE - the baseline setting resulted in the highest TIF mode, followed by the a- and c-prior setting, and last by the c-prior only setting. Even though the ordering depended on the particular program, the baseline and c-prior only settings were generally most different from one another.

Conclusions

The current study compared item calibration results from four IRT software programs, using four model combinations and three sets of item priors. More specifically, comparisons were made between software programs using settings that were as similar as possible and within each program using three different item prior settings. Overall, fewer between-program and within-program differences were found when the 2PL model was used in comparison to when the 3PL model was used. However, model selection for the FR items appeared to have little

impact on item estimates, TCCs, and TIFs. Therefore, in situations where it may be required to adopt a new software program or alter current calibration settings, choosing a model with fewer parameters is likely to result in fewer ramifications. This recommendation is made with the assumption that the more simplistic model adequately fits the data.

In general, it appeared as though the results from MULTILOG (as opposed to PARSCALE) were more consistent with IRTPRO. This was expected since IRTPRO was designed with the intent to replace MULTILOG, but even so, this finding is encouraging. Therefore, it may reasonable to expect fewer discrepancies when moving from MULTILOG to IRTPRO than moving from PARSCALE to IRTPRO.

This study provided a comprehensive comparison between two well-established software programs and two recently developed programs. However, it is important to stress that the results presented were for one test form only and should be replicated using different subject areas in order to gain a better understanding of the similarities or differences between programs. Furthermore, the current study also investigated whether differences at the item level carried over to meaningful differences at the test level by first looking at item parameter estimates followed by TCCs and TIFs. However, it is also important to expand comparisons made in this study to contexts that are more aligned with operational settings. Typically within a testing program, there are multiple forms of an exam that in turn, require the use of equating methods. Therefore, it is important to understand the extent to which the differences found in this study can potentially affect estimated equating relationships. As a result, the next study presented is an extension of the current study and compares IRT true score equating relationships and AP grade agreements using the same data used in the current study.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Cai, L. (2012). *flexMIRT* (Version 1.0) [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Thissen, D. J., & du Toit, S. (2011). *IRTPRO* (Version 2.1) [Computer program]. Mooresville, IN: Scientific Software.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351-363.

Muraki, E., & Bock, R. D. (2003). *PARSCALE* (Version 4.1) [Computer program]. Mooresville, IN: Scientific Software.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (pp. 43-75). New York, NY: Taylor & Francis.

Thissen, D. J., Chen, W.-H., & Bock, R. D. (2003). *MULTILOG* (Version 7.0) [Computer program]. Mooresville, IN: Scientific Software.

Toit, M. D. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.

Table 1

Prior Distributions Used During Item Parameter Estimation

Condition	Slope (a)	Location (b)	Pseudo-guessing (c)
Baseline	Normal (1.133, 0.604)	None	Logit Normal (-1.39, 0.5)
a- and c- prior	Normal (1.0, 1.0)	None	Logit Normal (-1.39, 0.5)
c-prior only	None	None	Logit Normal (-1.39, 0.5)

Note. PARSCALE used *Beta* (5, 20) on the *c*-prior for all conditions.

Table 2

Calibration Settings for Item Parameter Estimation Procedures for All Conditions

Setting	MULTILOG	PARSCALE	IRTPRO	flexMIRT
# of cycles for E steps	3,000	3,000	3,000	3,000
# of cycles for M steps	3,000	3,000	3,000	3,000
Convergence criterion for E steps	0.0001	0.0001	0.0001	0.0001
Convergence criterion for M steps	0.0001	0.0001	0.0001	0.0001
# of quadrature points	49	49	49	49
Quadrature Range	[-6, 6]	[-6, 6]	[-6, 6]	[-6, 6]

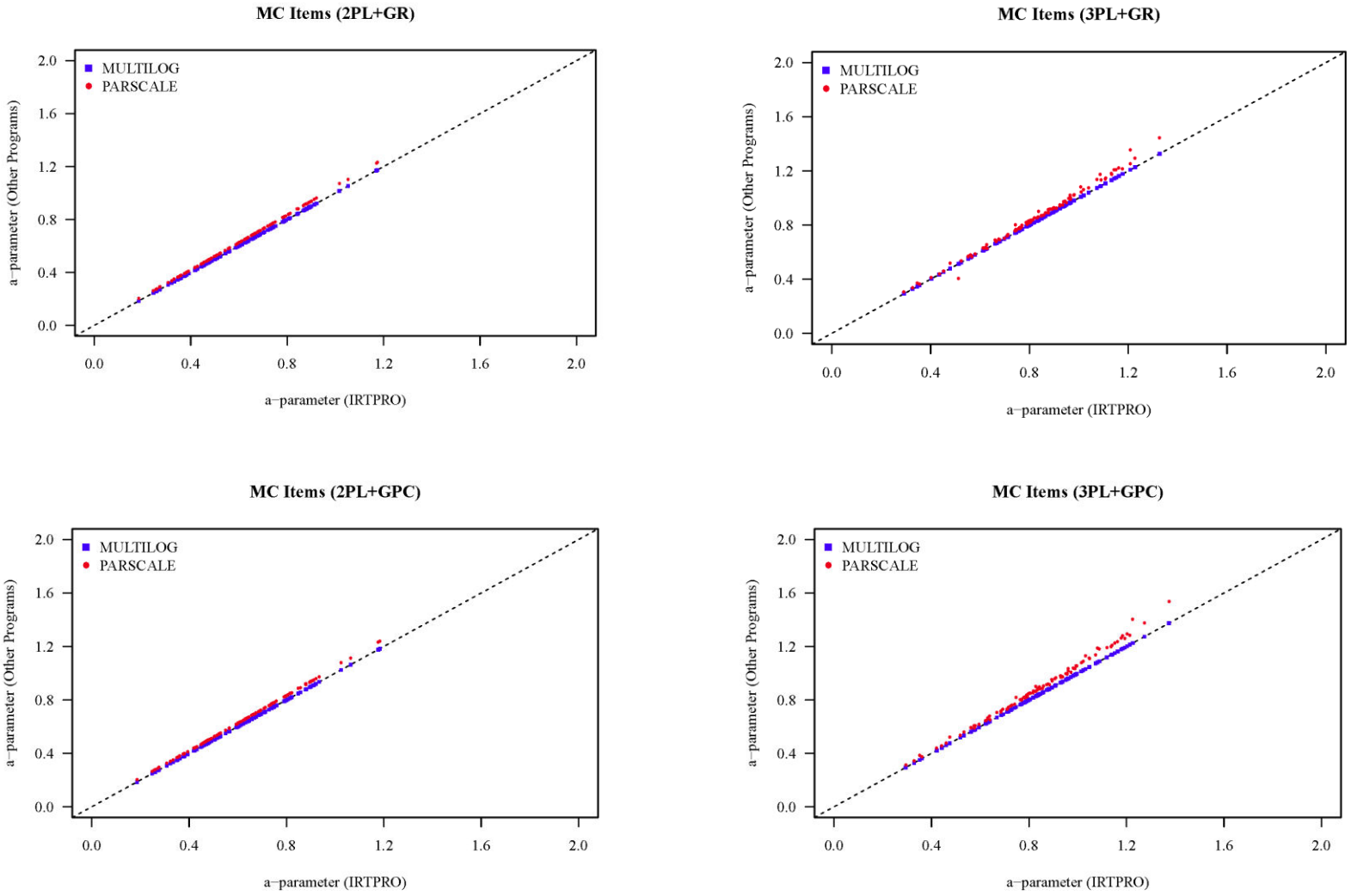


Figure 1. Between-program comparisons of MC slope estimates using the Baseline setting.

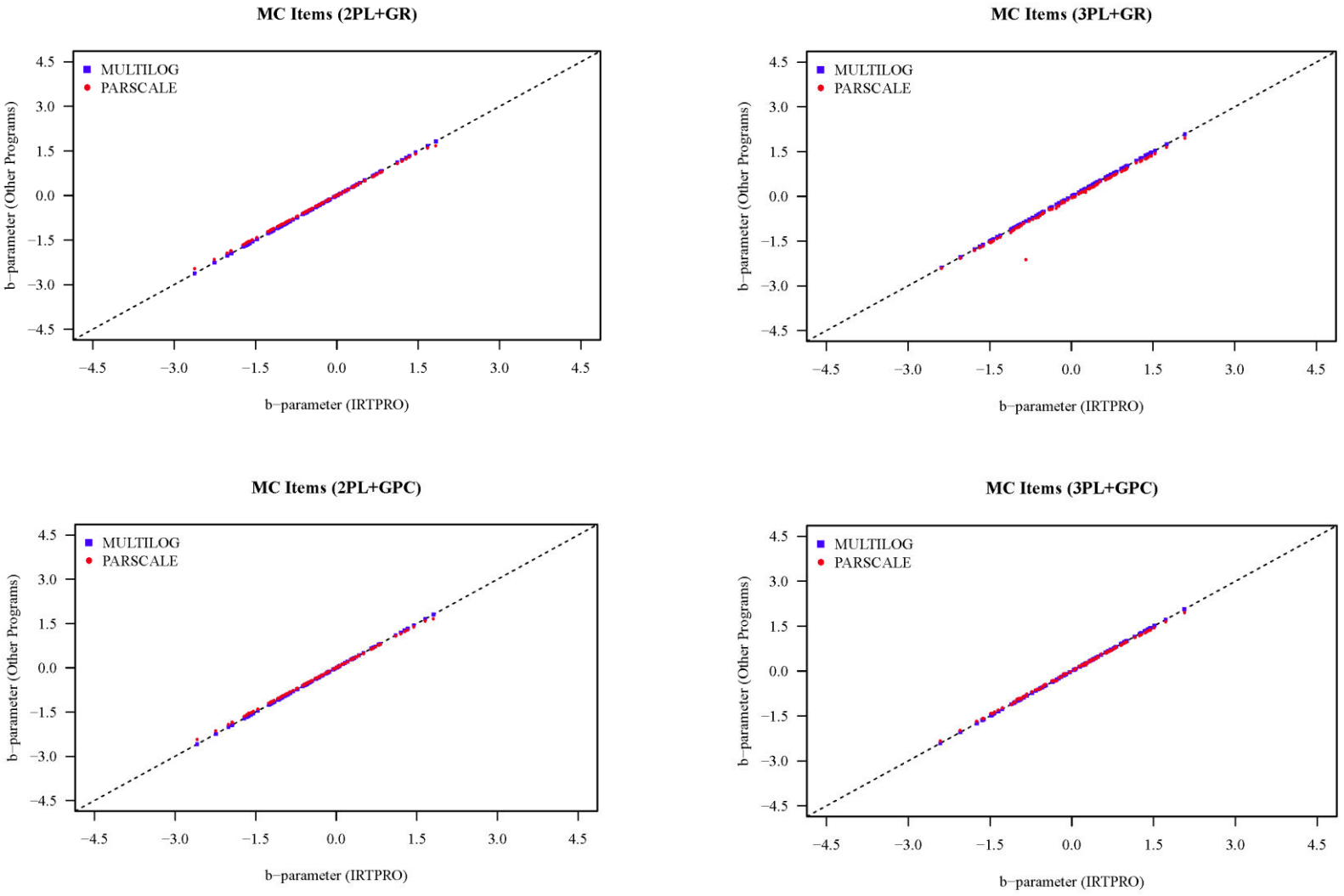


Figure 2. Between-program comparisons of MC location estimates using the Baseline setting.

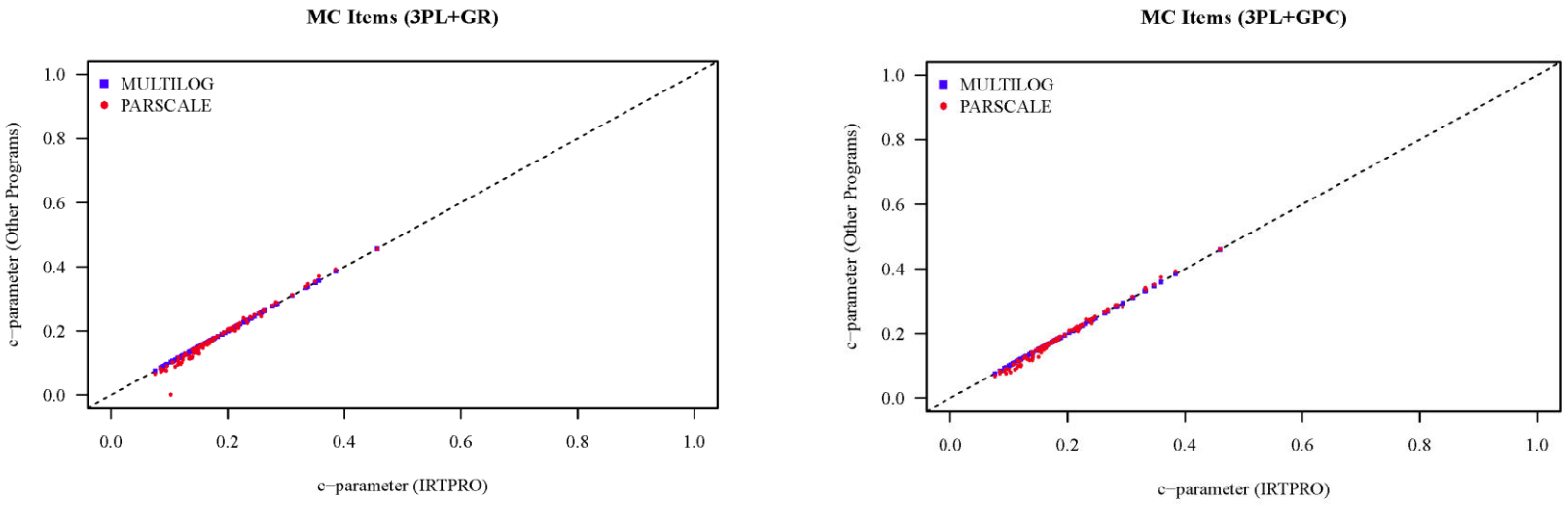


Figure 3. Between-program comparisons of MC pseudo-guessing estimates using the Baseline setting.

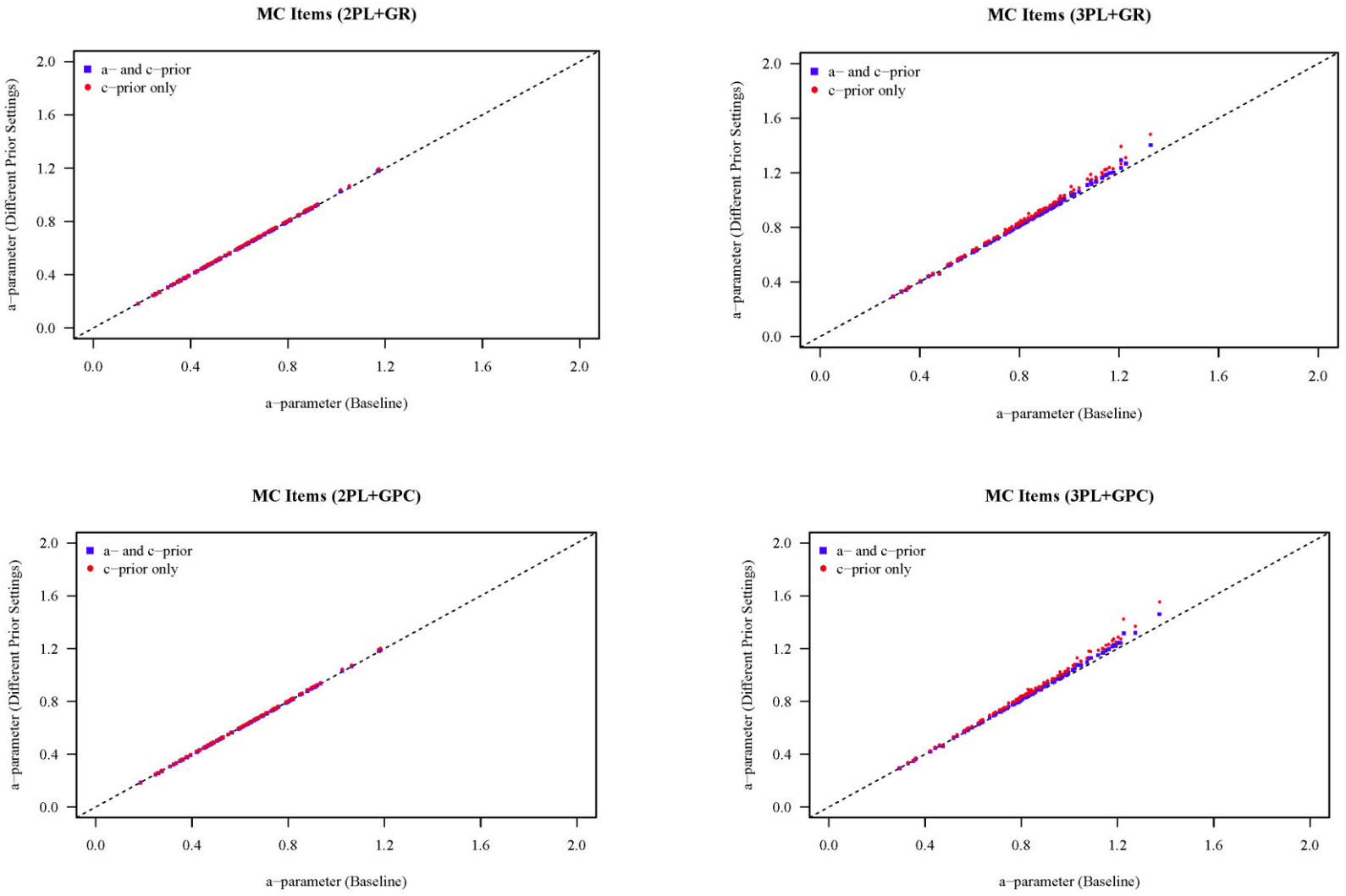


Figure 4. Comparison of MC slope estimates using each item prior setting in IRTPRO.

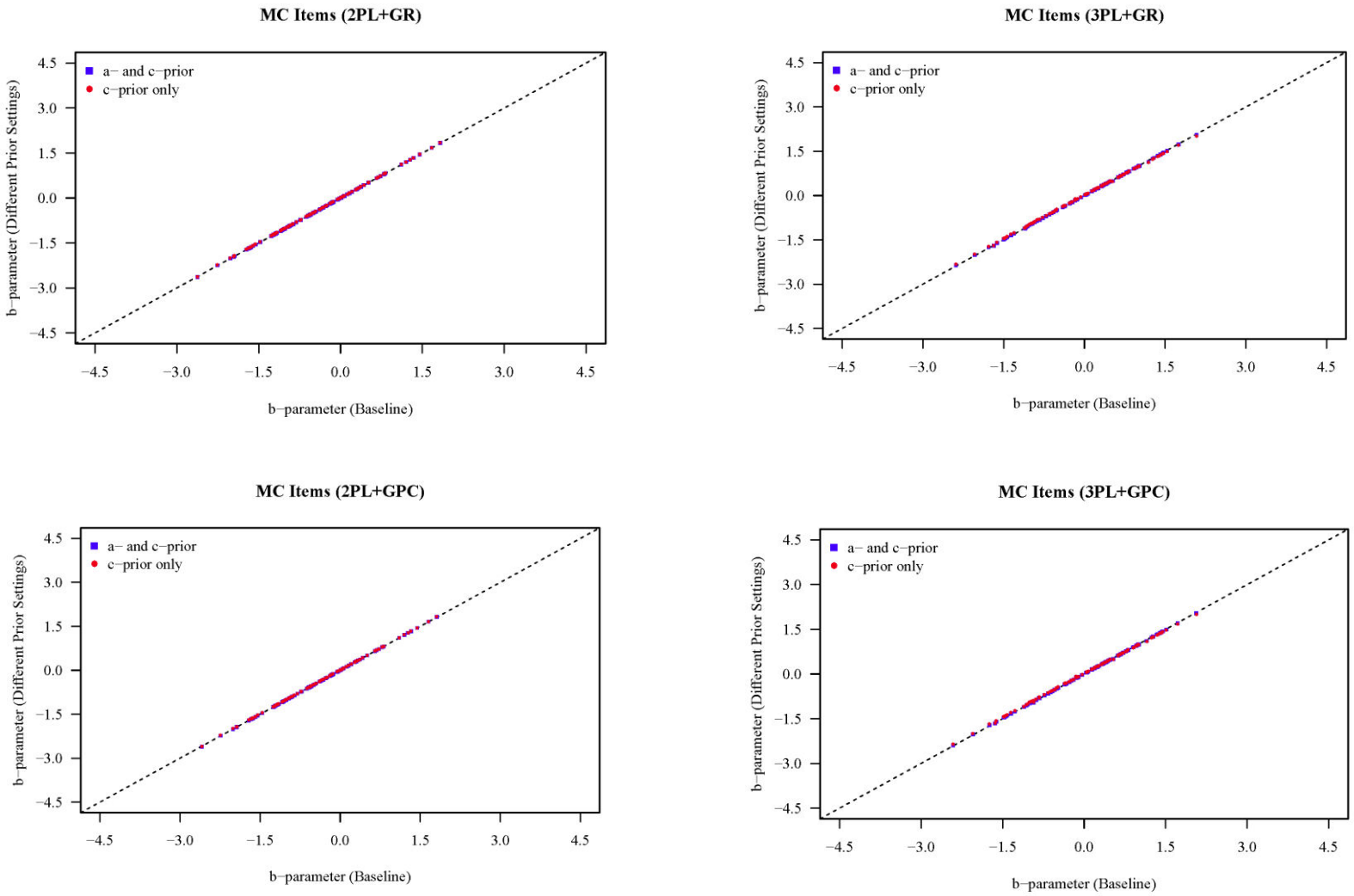


Figure 5. Comparison of MC location estimates using each item prior setting in IRTPRO.

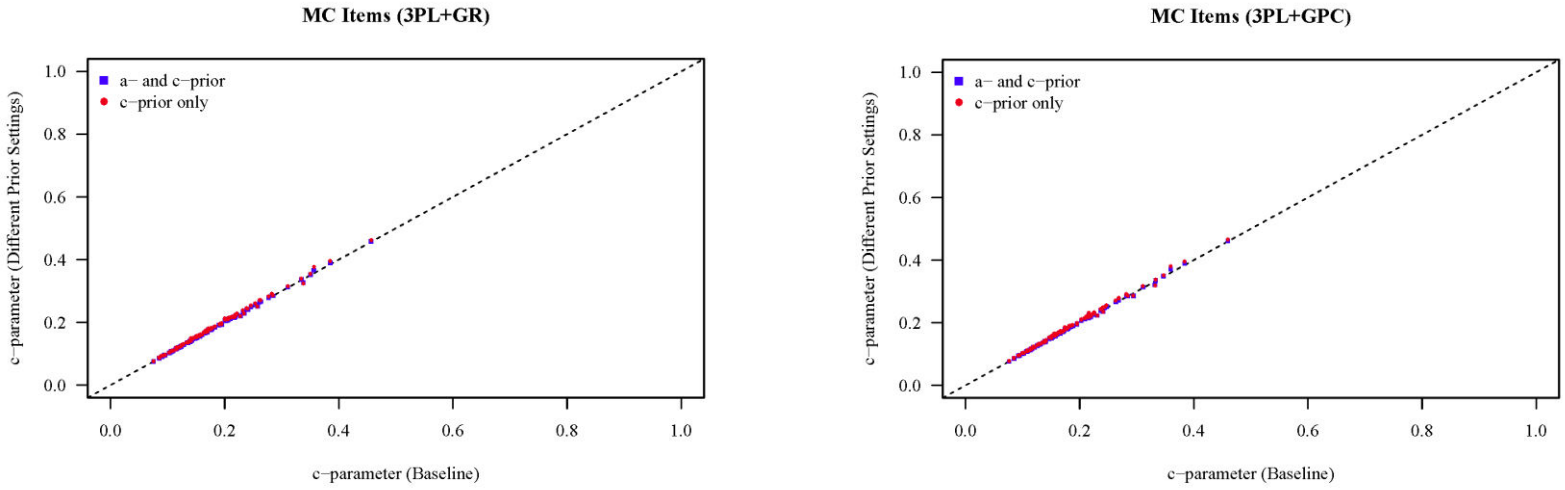


Figure 6. Comparison of MC pseudo-guessing estimates using each item prior setting in IRTPRO.

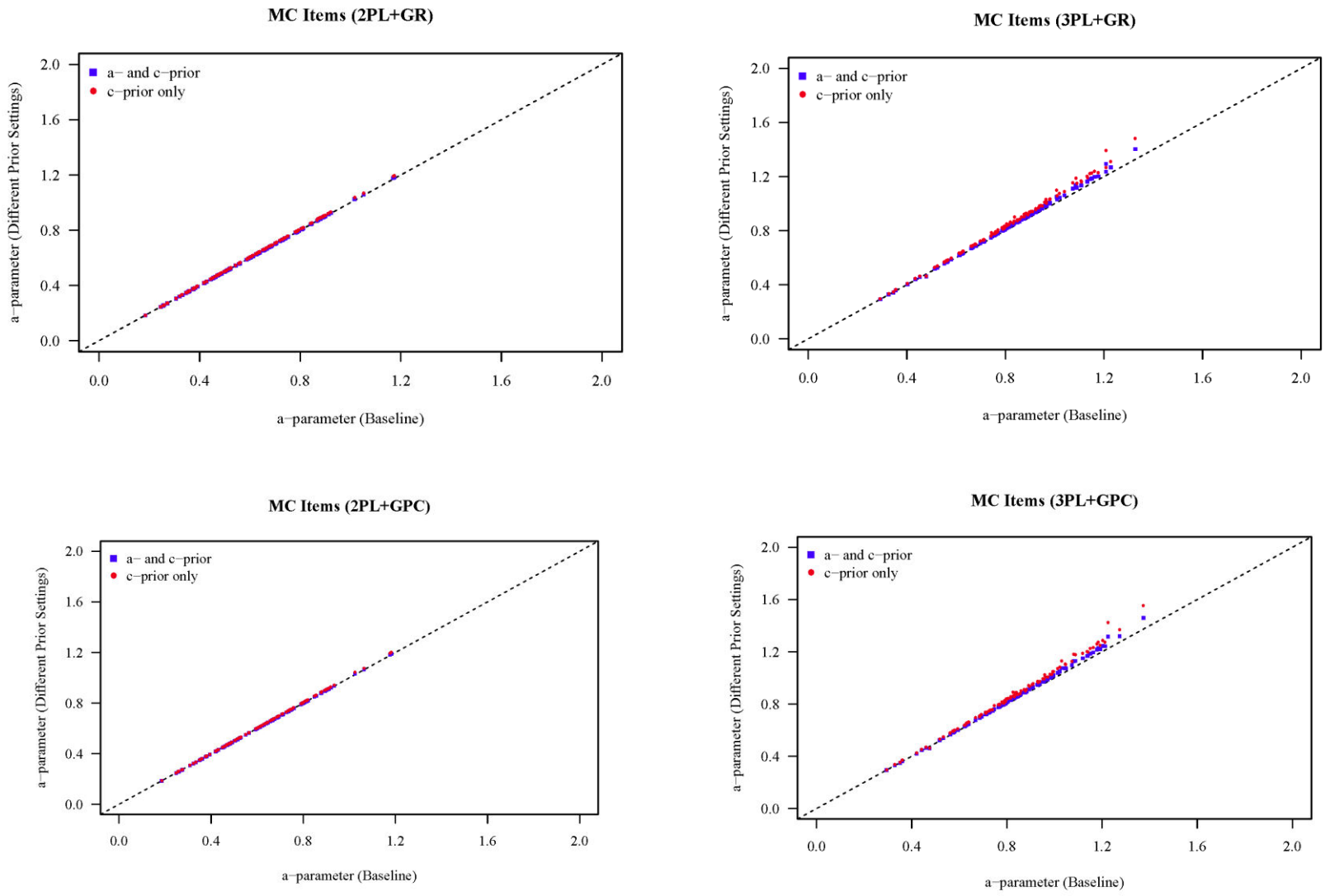


Figure 7. Comparison of MC slope estimates using each item prior setting in MULTILOG.

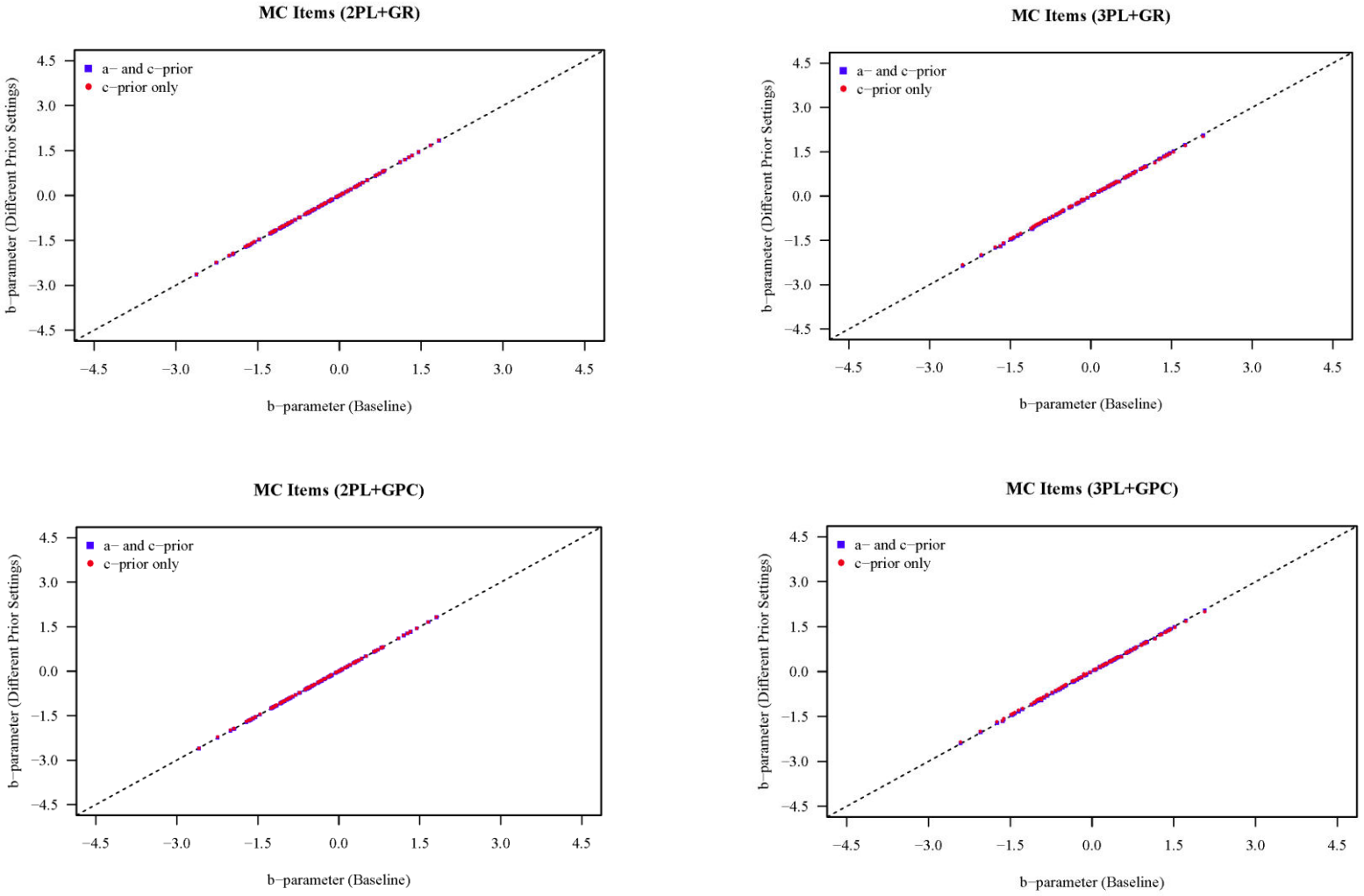


Figure 8. Comparison of MC location estimates using each item prior setting in MULTILOG.

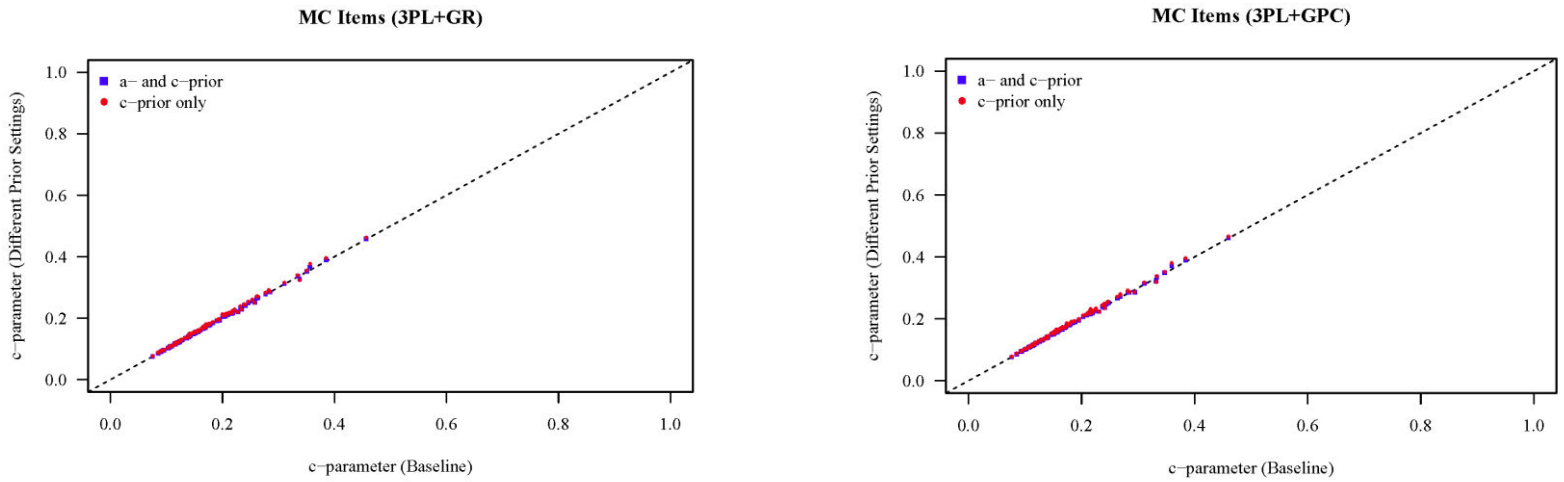


Figure 9. Comparison of MC pseudo-guessing estimates using each item prior setting in MULTILOG.

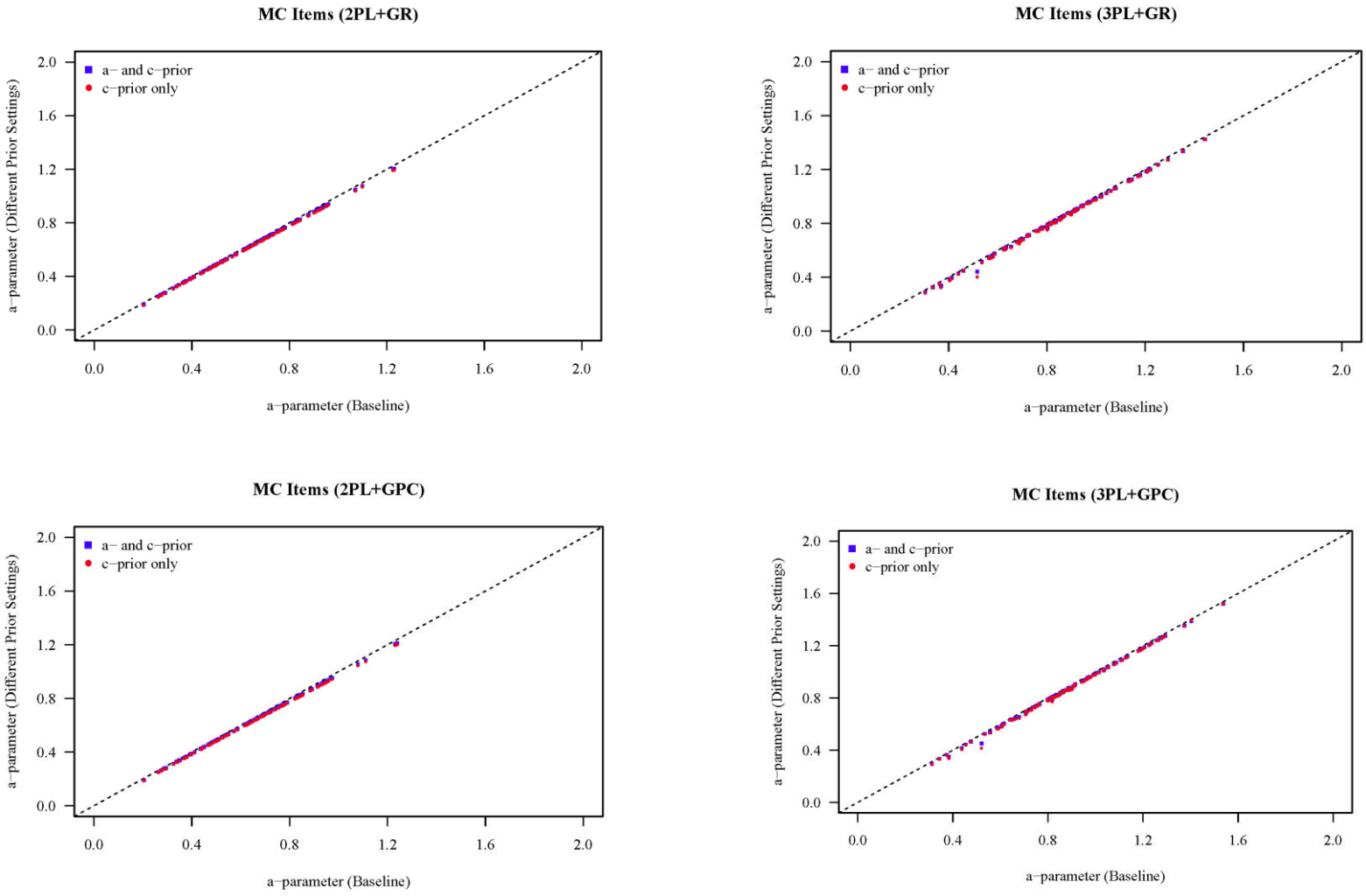


Figure 10. Comparison of MC slope estimates using each item prior setting in PARSCALE.

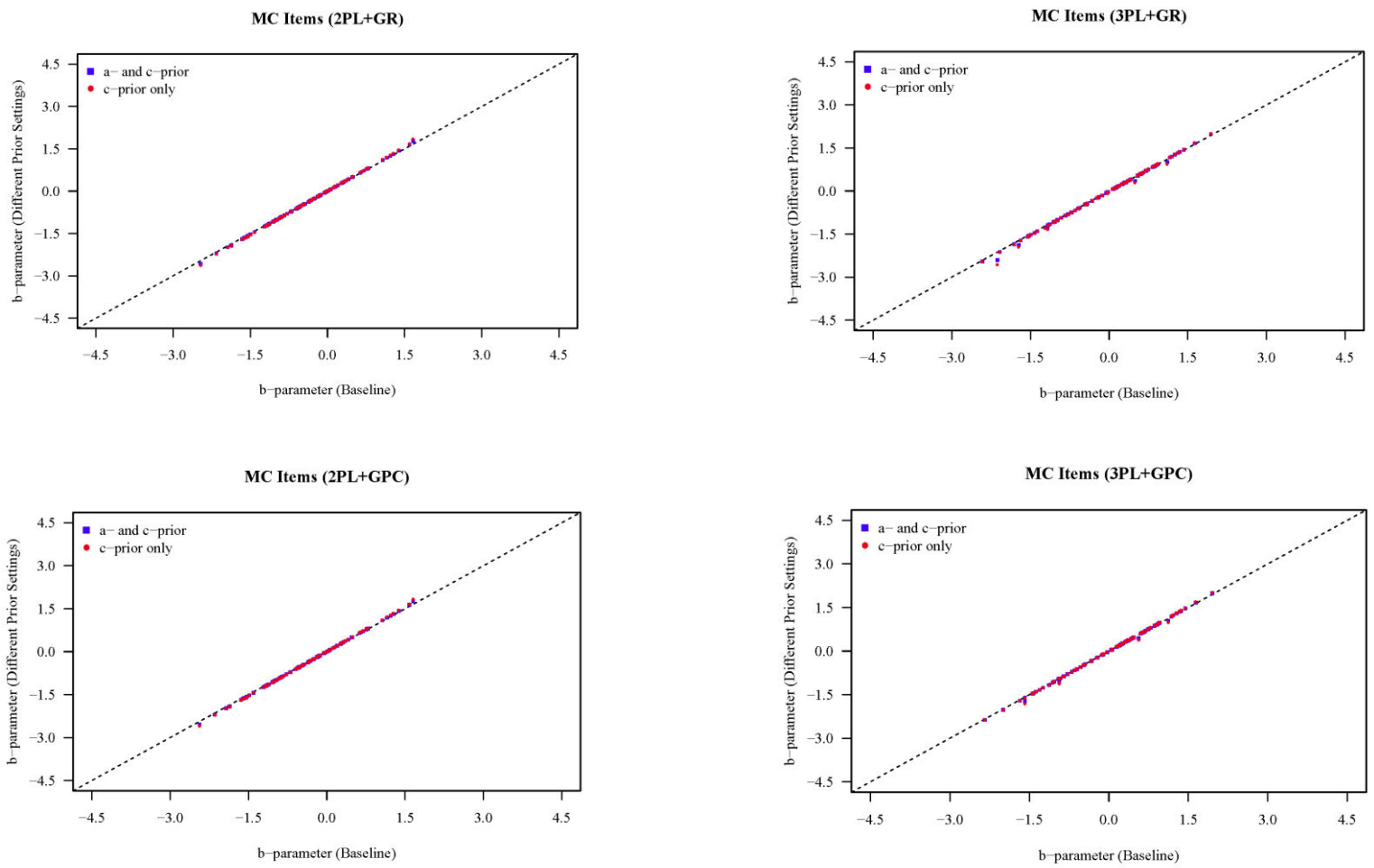


Figure 11. Comparison of MC location estimates using each item prior setting in PARSCALE.

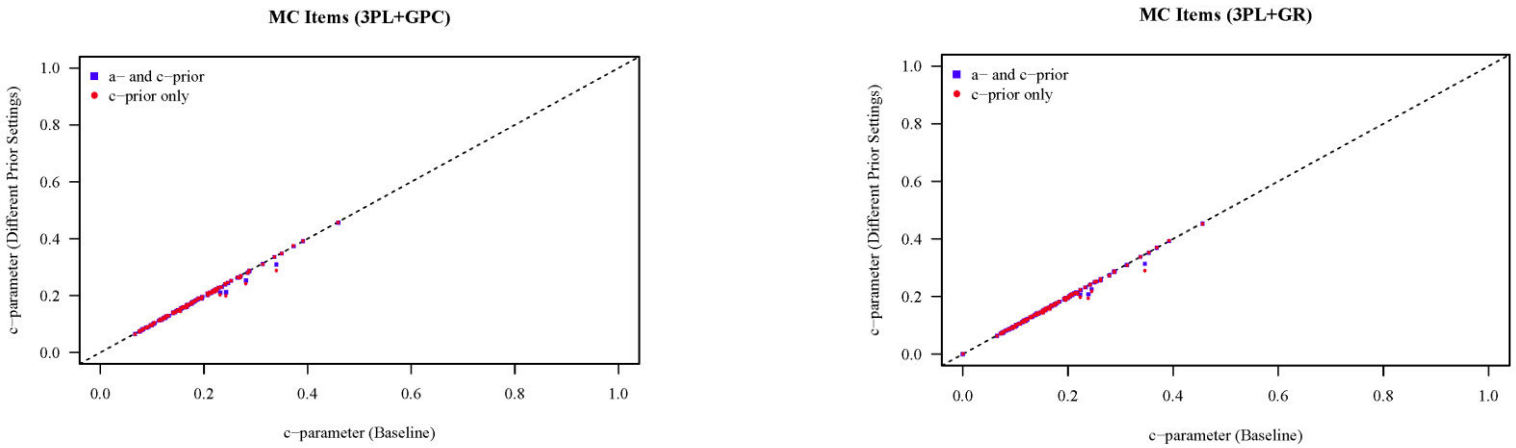


Figure 12. Comparison of MC pseudo-guessing estimates using each item prior setting in PARSCALE.

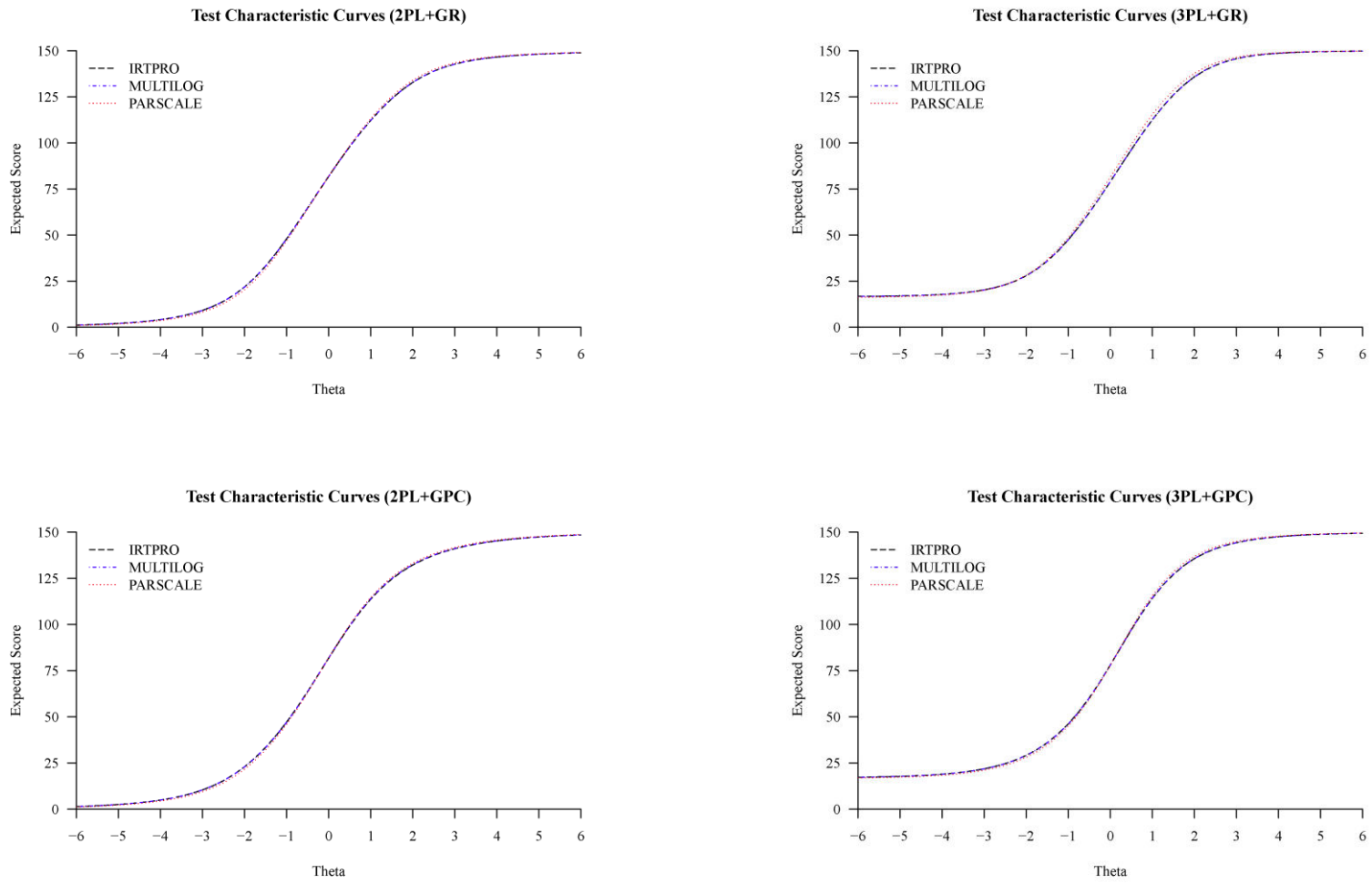


Figure 13. Between-program comparisons of TCCs using the Baseline setting.

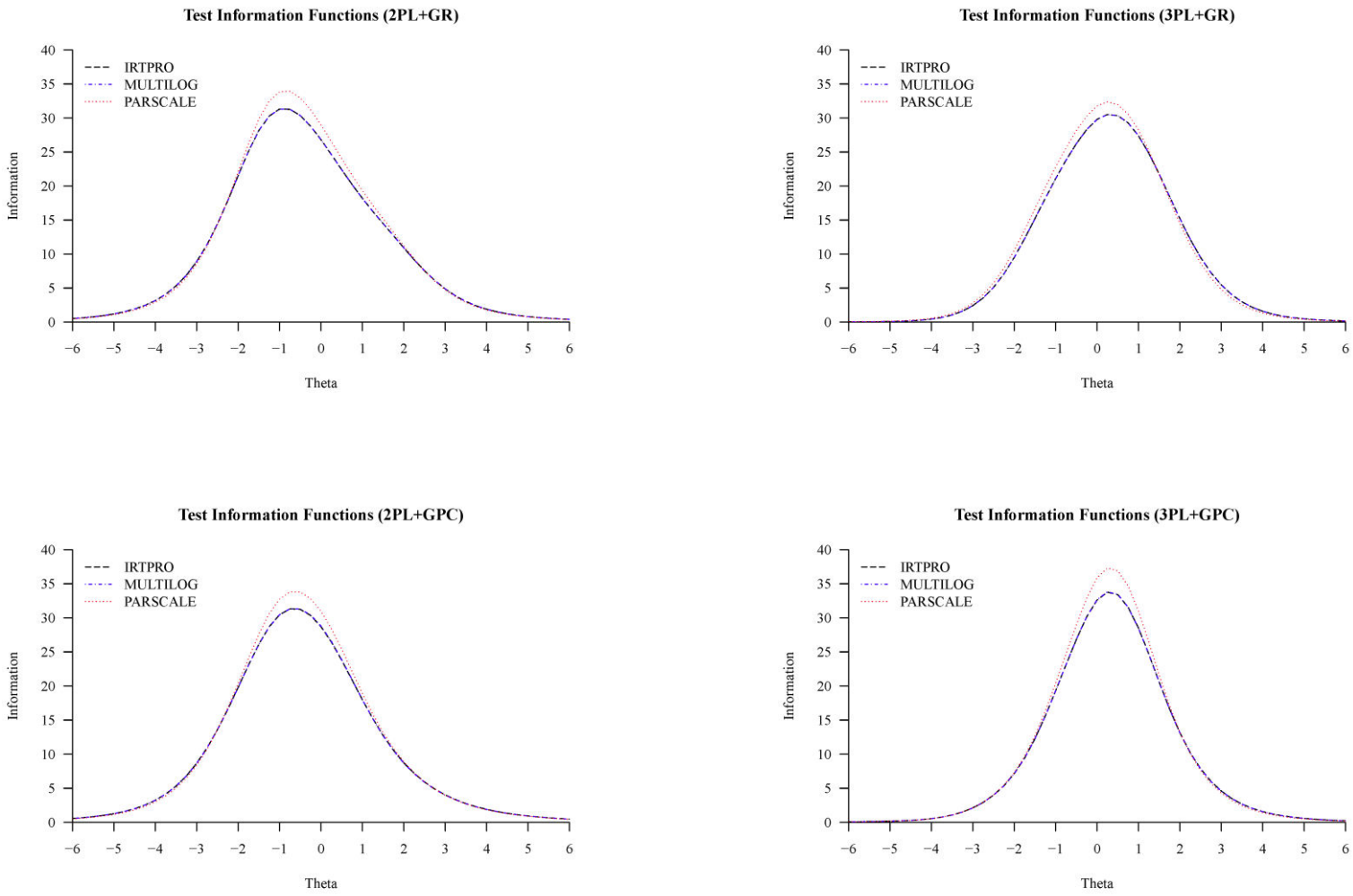


Figure 14. Between-program comparisons of TIFs using the Baseline setting.

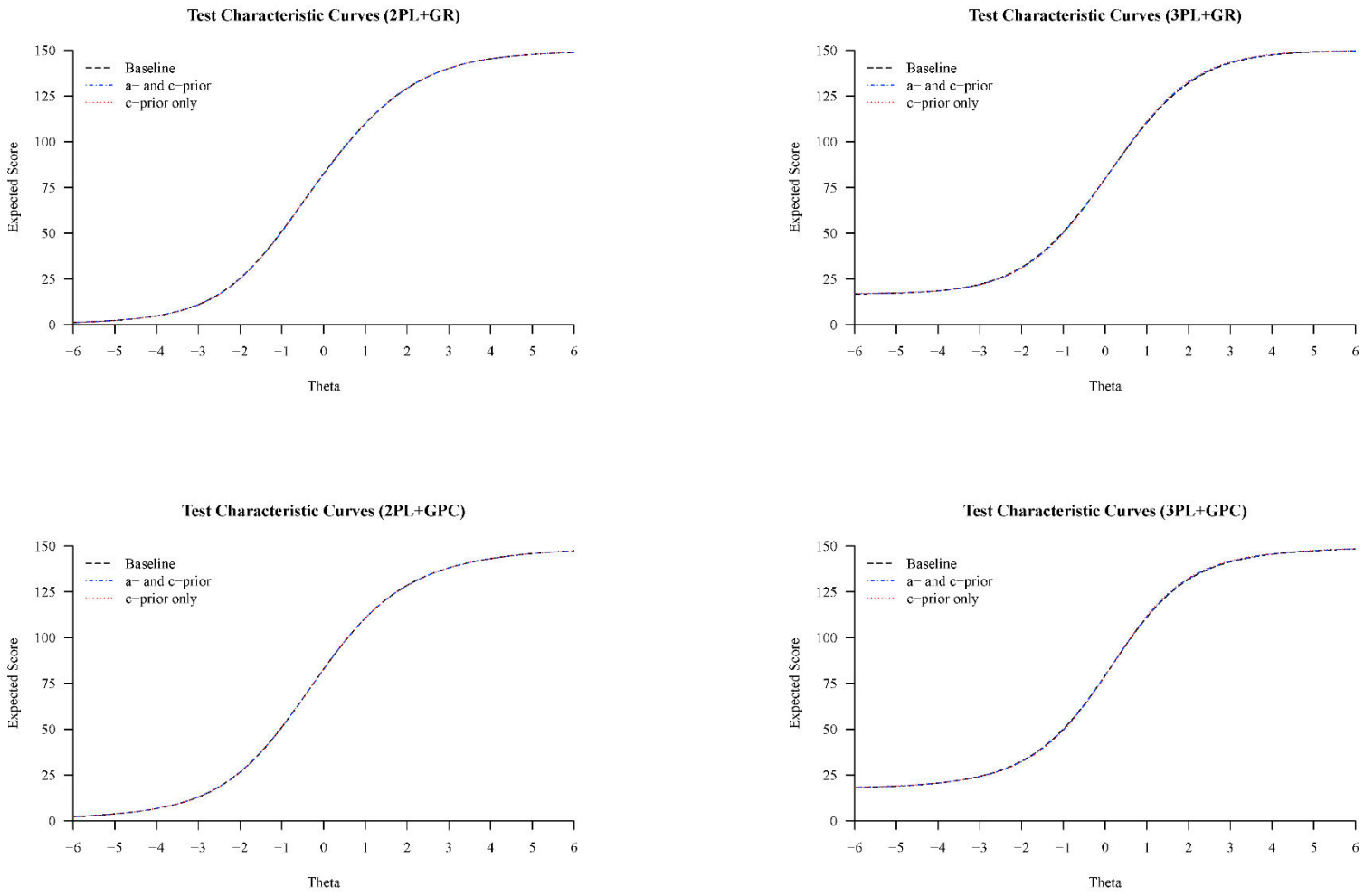


Figure 15. Comparison of TCCs using each item prior setting in IRTPRO.

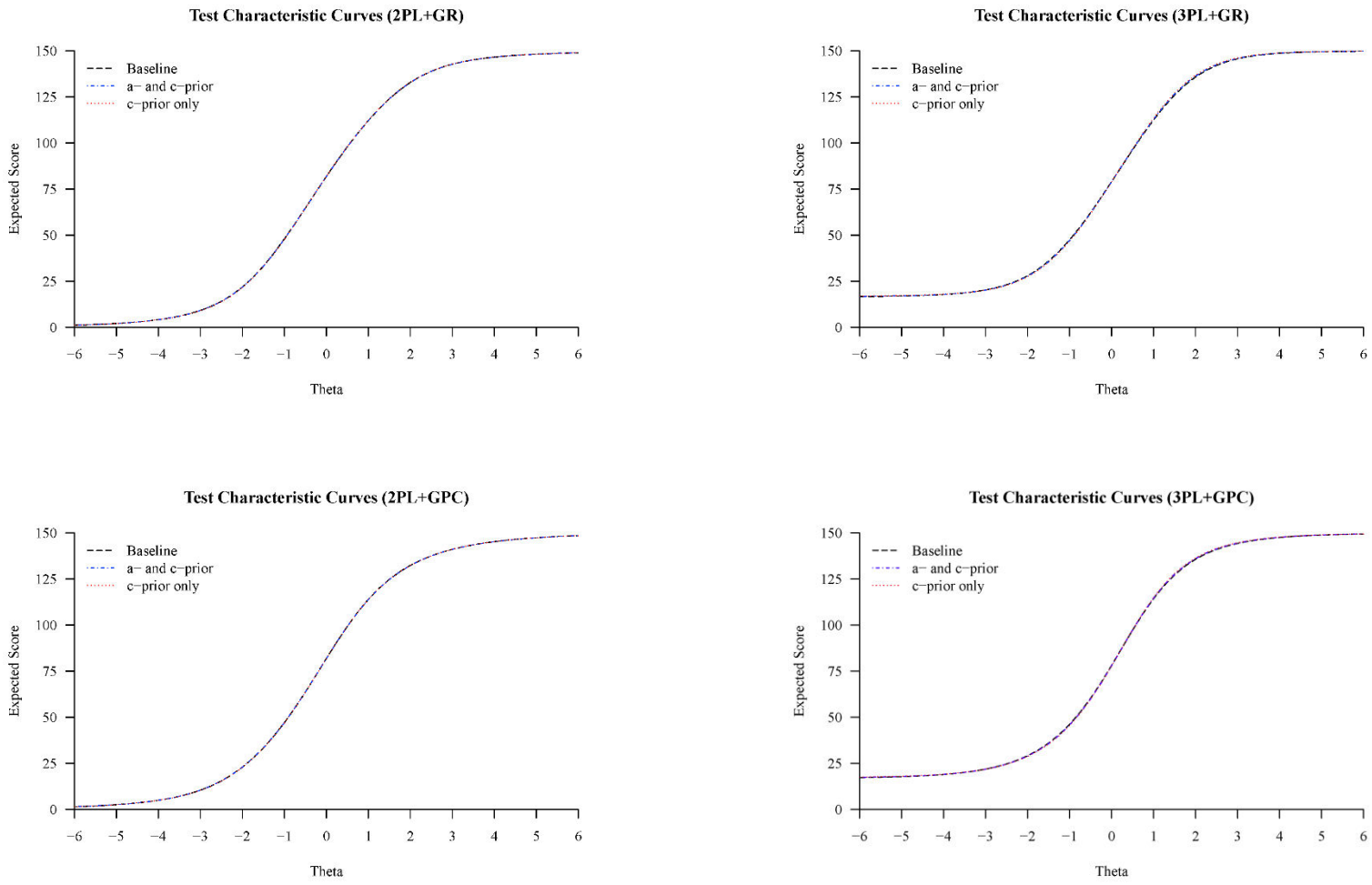


Figure 16. Comparison of TCCs using each item prior setting in MULTILOG.

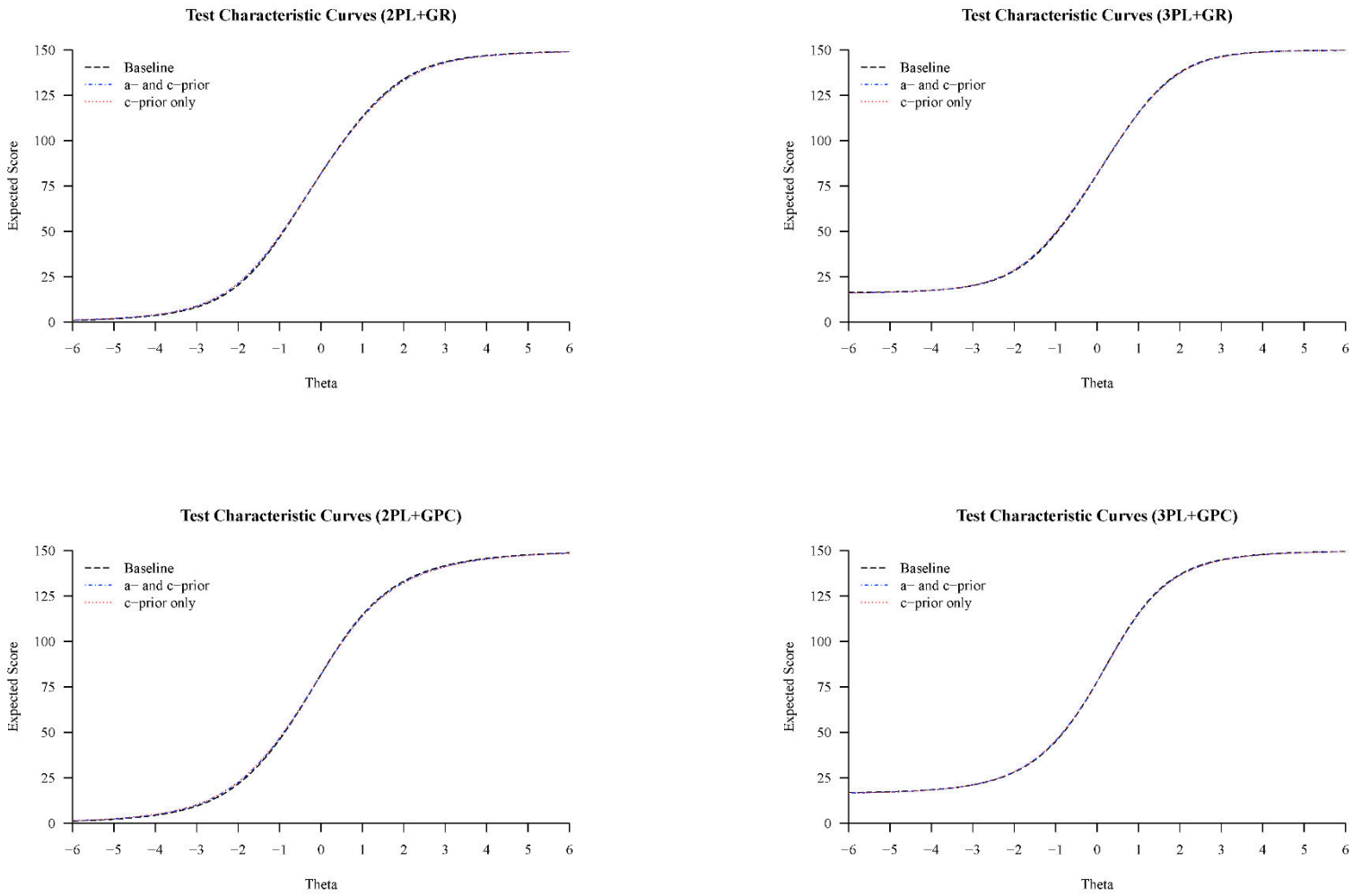


Figure 17. Comparison of TCCs using each item prior setting in PARSCALE.

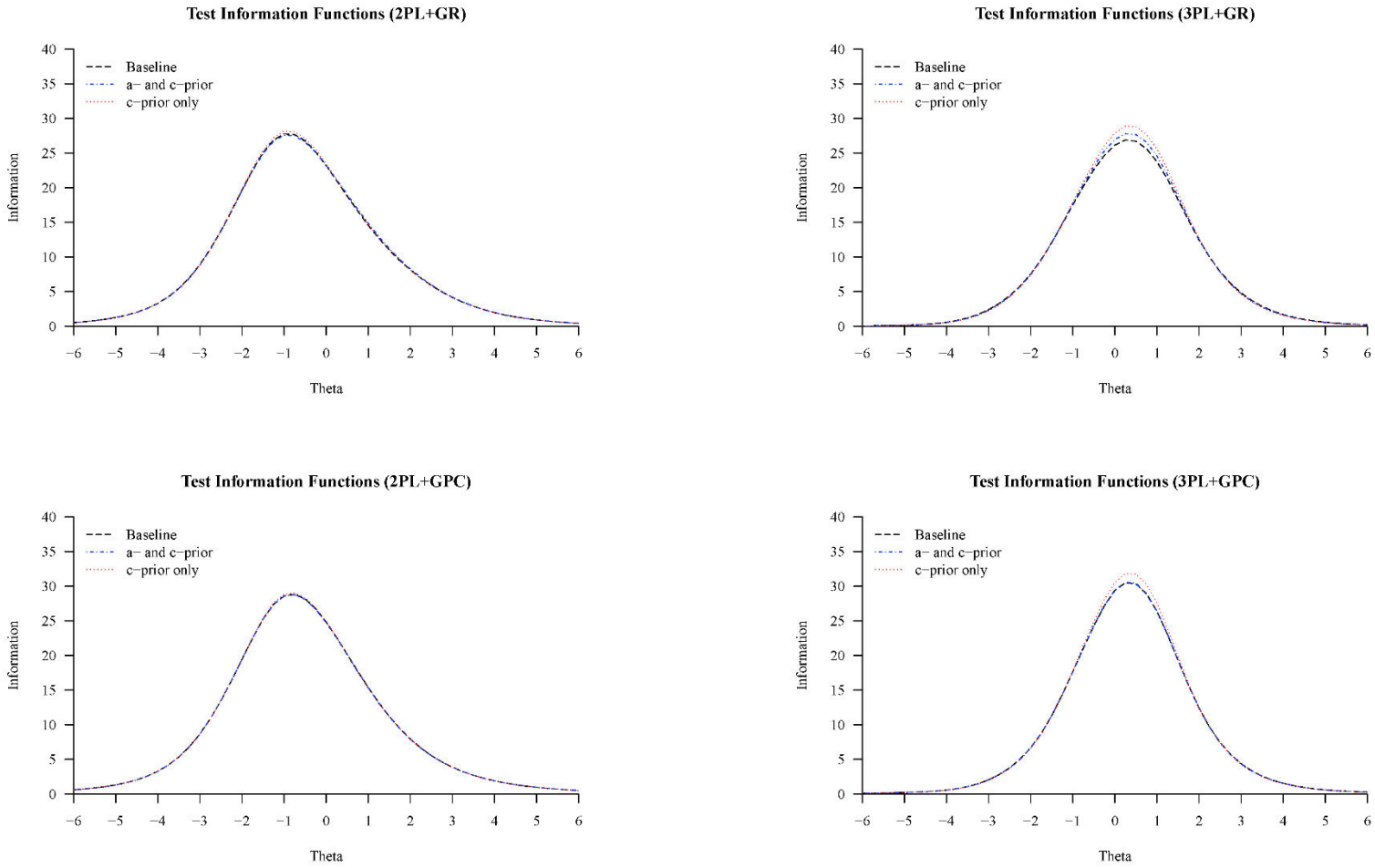


Figure 18. Comparison of TIFs using each item prior setting in IRTPRO.

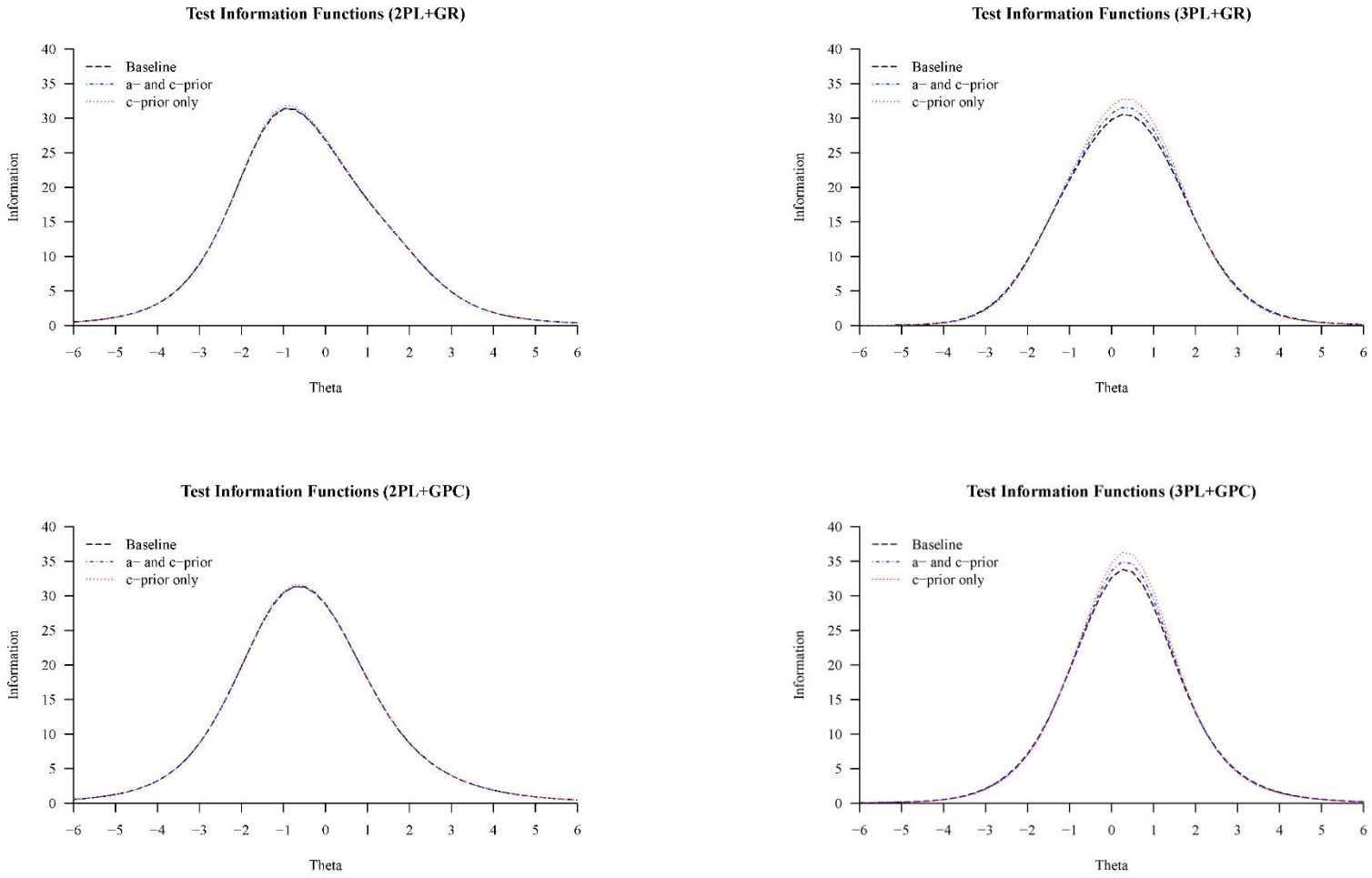


Figure 19. Comparison of TIFs using each item prior setting in MULTILOG.

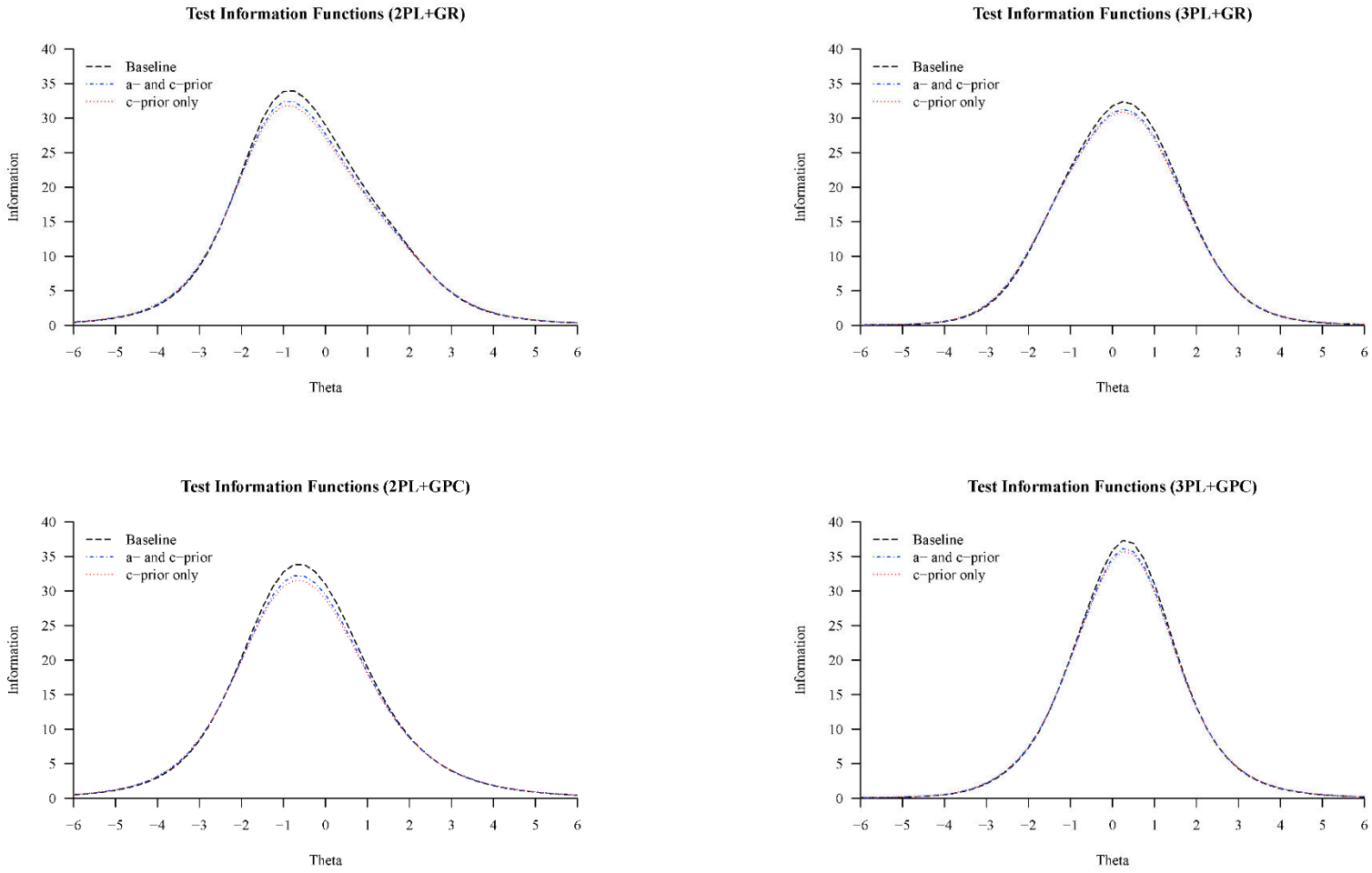


Figure 20. Comparison of TIFs using each item prior setting in PARSCALE.

Appendix

Steps to compute category and item information functions for the GR model (de Ayala, 2009):

1) Compute the cumulative probabilities P_0^* to P_{m+1}^* as:

$P_0^* = 1$,

$P_{m+1}^* = 0$, and for all others,

$$P_{xj}^* = \frac{exp^{a_j(\theta - b_{xj})}}{1 + exp^{a_j(\theta - b_{xj})}},$$

where a_j is the discrimination for item *j* and b_{xj} is the threshold of category *x* of item *j*.

2) Compute the probability of responding in a specific category as:

$$P_{jk} = P_{jk}^* - P_{jk+1}^*$$

$$P_0 = P_0^* - P_1^*$$

$$P_1 = P_1^* - P_2^*$$

... ...

$$P_{10} = P_{10}^*$$

3) Compute 1st derivative of each cumulative probability function as:

$$P_0' = 0 - a_j \left[\frac{exp^{a_j(\theta - b_1)}}{\left(1 + exp^{a_j(\theta - b_1)}\right)^2} \right]$$

$$P_1' = a_j \left[\left(\frac{exp^{a_j(\theta - b_1)}}{\left(1 + exp^{a_j(\theta - b_1)}\right)^2} \right) - \left(\frac{exp^{a_j(\theta - b_2)}}{\left(1 + exp^{a_j(\theta - b_2)}\right)^2} \right) \right]$$

... ...

$$P_{10}' = a_j \left[\frac{exp^{a_j(\theta - b_{10})}}{\left(1 + exp^{a_j(\theta - b_{10})}\right)^2} \right]$$

4) Compute the category information functions as:

$$I_{xj}(\theta) = \frac{\left(P_{xj}'\right)^2}{P_{xj}}.$$

5) Compute item information as:

$$I_j(\theta) = \sum_{xj=0}^{m} \frac{\left(P_{xj}'\right)^2}{P_{xj}}.$$

Steps to compute category and item information functions for the GPC model (Muraki, 1993):

1) Compute the probability of endorsing each category as:

$$P_{ijk}(\theta) = \frac{exp^{\left[\sum_{h=1}^{k} a_j(\theta_i - b_j + d_{jh})\right]}}{\sum_{g=1}^{m_j} exp^{\left[\sum_{h=1}^{g} a_j(\theta_i - b_j + d_{jh})\right]}},$$

where b_j is the difficulty for item *j*, d_{jh} is the threshold boundary *h* of item *j*, and *g* is an index used in the summation of the denominator over the *m* categories.

1a) For Bock's nominal model (used in MULTILOG only) the probability of endorsing each category is computed as,

$$P_{ijk}(\theta) = \frac{exp(a_{jk}\theta + c_{jk})}{\sum_{h=1}^{m_j} exp(a_{jh}\theta + c_{jh})},$$

where a_{jk} and c_{jk} are the slope and intercept parameters, respectively, associated with the k^{th} category of item *j*, and *h* is used to sum over the *m* categories in the denominator. The remaining steps apply to both the GPC and nominal models.

2) Compute $\bar{T}_j\ (\theta)$, for *h* = 1 to *k* +1. (For an item scored 0-10, *k* = 10)

$$\bar{T}_j\ (\theta) = \sum_{h=1}^{k} T_h P_{jh}(\theta),$$

where T_h is the score function or observed score for item *j*.

3) Compute the item category information function as,

$$I_{jk}(\theta) = a_j^2\ \left[\left(T_h - \bar{T}_j\right)^2 P_{ijk}(\theta)\right].$$

4) Compute the item information function as,

$$I_j(\theta) = \sum_{h=1}^{k} I_{jk}(\theta).$$

Chapter 5: A Comparison of Several Item Response Theory Software Programs with Implications for Equating Mixed-Format Exams

Jaime Peterson, Mengyao Zhang, Shichao Wang, Seohong Pak,
Won-Chan Lee, and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

The current study is a direct extension of Chapter 4 and empirically investigated effects associated with the use of different IRT calibration programs and item prior settings on estimated IRT true score equating relationships. Equating relationships were estimated using three different item prior settings, and in four calibration programs: MULTILOG, PARSCALE, IRTPRO and flexMIRT. Equating results were compared for raw composite scores, unrounded scale scores, and AP Grades. Using data from the 2011 AP Biology main and alternate forms, it was found that the equating results obtained from IRTPRO and flexMIRT were basically identical across all conditions. Furthermore, equating results between IRTPRO/flexMIRT and MULTILOG were also very similar, which supports the replacement of MULTILOG with either program. In general, when differences were present, they were generally found in conditions that included the 3PL model. Importantly, the small differences in equating results between programs did not carry over into AP Grades, and therefore examinees would have received the same Grades regardless of which calibration program was used.

A Comparison of Several Item Response Theory Software Programs with Implications for Equating Mixed-Format Exams

Equating is often necessary in operational testing programs that regularly administer multiple forms of a test. In recent years, more tests have transitioned from multiple choice (MC) only to mixed-format, typically meaning that they consist of MC and essay or free response (FR) items. Mixed-format tests are advantageous in that they can assess a broader range of cognitive skills more easily, but they can also complicate equating procedures. For example, when item response theory (IRT) equating procedures are used, distinct models are required for each item type. Furthermore, FR items have more response options than MC items and therefore require more parameters than dichotomously scored items. Naturally, models with more parameters are more prone to estimation error, which can then influence estimated equating relationships. Therefore, it is important to understand how the use of different IRT calibration programs and different item prior settings could carry over into estimated IRT equating relationships for mixed-format tests.

The current study is a direct extension of the study presented in Chapter 4, and compares estimated IRT true score equating relationships resulting from the use of different IRT calibration programs and several item prior settings. In that study, item parameter estimates, test information functions (TIFs), and test characteristic curves (TCCs) were found to differ by varying amounts as a result of using different calibration programs (Peterson et al., 2014). However, from that study, it was difficult to predict how differences might affect estimated equating relationships. For example, small differences in item parameter estimates may accumulate and produce different estimated equating relationships. On the other hand, differences may be too small to have an effect on the equating relationships, or they may cancel one another out, thereby having no effect on the equating relationships. Yet another possibility is that differences between software programs could manifest similarly for the Old and New forms, thereby resulting in the similar equating relationships across programs. With so many possibilities, it was important to empirically investigate the actual estimated equating relationships that resulted from using different IRT calibration programs and different item prior settings.

The findings in Peterson et al. (2014) suggest that going from one software program to another and/or making slight modifications to item prior settings tends to make more of a

difference when more complex models are used. More specifically, greater differences were found when the 3PL model was used for the MC portion of the exam; whereas, when the 2PL model was used, very few differences, if any, were detected. However, the pattern of differences was not always consistent across software programs and in some instances were in direct opposition to one another. Therefore, it was of interest to see whether these differences carried over into the IRT true score equating relationships.

The primary focus of this study was to evaluate differences in IRT true score equating relationships stemming from the use of different IRT software programs (differences related to item prior settings were secondary). This issue has become more central in the past few years with the introduction of two software programs, flexMIRT (Cai, 2012) and IRTPRO (Cai, Thissen & du Toit, 2011). The current study compares results provided by these two newer software programs to those provided by two well-established programs that have been researched more extensively – MULTILOG (Thissen, Chen, & Bock, 2003) and PARSCALE (Muraki & Bock, 2003). It was especially important to compare MULTILOG and IRTPRO since the latter was designed to replace the former. Furthermore, the comparison of IRTPRO with flexMIRT was also of interest due to the overlap in authors and theoretical frameworks between the two programs.

Background Information

Since the current study was a direct extension of the Peterson et al. (2014) study, presented in Chapter 4, details concerning program-specific transformations and IRT model components are not provided. Instead, an overview of the IRT model combinations and IRT software is provided, and interested readers have the option to refer to Chapter 4 for specific details. Since this is an extension study, the item calibration results presented in Chapter 4 are directly applicable to the current study and were actually used to conduct IRT true score equating. The major difference between the study by Peterson et al. (2014) and the current study is the necessary inclusion of the alternate or new form examinees in the latter. In Chapter 4, analyses were presented for the main form sample only, because it was determined that findings were consistent across main and alternate form samples. Even though results were not presented for the alternate form sample, they were carried out and were used in the current study to conduct IRT true score equating.

IRT Model Combinations

Since the exams used in this study were mixed-format, different IRT models were fit to the multiple choice (MC) and free response (FR) items. For the MC items, the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) model were examined. For the FR items, Samejima's (1969) Graded Response (GR) model and Muraki's (1992) Generalized Partial Credit (GPC) model were used. In total, four model combinations were used: 2PL + GR, 3PL + GR, 2PL + GPC, and 3PL + GPC.

IRT Calibration Programs and Settings

Four IRT calibration programs were used to estimate item parameters under each of the four IRT model combinations: MULTILOG, PARSCALE, IRTPRO, and flexMIRT. This enabled the comparison between two well-established programs (i.e., MULTILOG and PARSCALE) and two newer programs (i.e., IRTPRO and flexMIRT).

Within each program, parameters for the four IRT model combinations were estimated using three different item prior settings, which resulted in each form being calibrated 12 times using the same program. In order to have a reference point, one of the three item prior settings was arbitrary chosen to serve as the baseline setting. The baseline setting specified item prior distributions on the *a*- and *c*-parameters (if present in the model). The remaining two item prior settings are referred to as the "a- and c-prior" setting and the "c-prior only" setting. All prior distributions settings, along with calibration settings, can be found in Tables 1 and 2, respectively.

As seen in Table 2, the general calibration settings were made as similar as possible across calibration programs in an attempt to reduce random noise. However, differences between programs were still possible because each was developed with different theoretical frameworks and by different authors (with the exception of IRTPRO and flexMIRT). Therefore, it is possible that different algorithms are used within each program, which could lead to small differences in results. Furthermore, it is possible for item prior settings to be used differently during the estimation processes within each program, which could influence results. It should be noted that the item prior distributions available in MULTILOG and PARSCALE were not fully compatible with each other and as a result an approximation was used for the *c*-parameter prior in PARSCALE. Specifically, the beta distribution used in PARSCALE has a very similar shape to the logit-normal distribution used in MULTILOG, IRTPRO, and flexMIRT.

Method

Data

The AP Biology main and alternate forms from 2011 were used in the current study, and can be regarded as the old and new forms, respectively. One obvious difference between the Peterson et al. (2014) study and the current study is that the latter included samples from both the main and alternate forms of the 2011 AP Biology exams. Data were collected operationally using the common item non-equivalent groups (CINEG) design. The main form consisted of 100 MC items with 5 options each, and 4 FR items. The MC items were dichotomously scored as 0 or 1, and FR items were scored by human raters on a 0-10 scale. The same scoring protocol was used with the alternate form, except there were 99 MC items and 4 FR items. Of the 100 and 99 MC items on the main and alternate forms, respectively, 24 were common to both forms. There were no common FR items. The main and alternate forms were administered operationally to 179,506, and 3,323 examinees, respectively. Due to sample size restrictions of MULTILOG, a random sample of 3,000 examinees from each form was drawn using SAS and used to conduct all analyses. Items on both forms were calibrated using four IRT model combinations, within four IRT software programs, and using three different calibration settings.

Linking and Equating Procedures

Because data were collected according to the CINEG design, scale linking was necessary in order to first put item and ability parameter estimates from both forms on the same scale before equating could be done. Using the MC common items, the Stocking and Lord (1983) test characteristic method was used to link the alternate and main form ability scales.

Next, IRT true score equating was performed to equate alternate form scores to main form equivalent scores. Operational non-integer weights were applied to alternate and main form composite scores. For the main form, the MC summed score was multiplied by a weight of 0.9 and the FR summed score was multiplied by a weight of 1.5 to form the composite score. These composite scores ranged from 0 to 150. For the alternate form, the MC summed score was multiplied by a weight of 0.909 and the FR summed score was multiplied by a weight of 1.5 to form the composite score. These composite scores also ranged from 0 to 150. IRT true score equating procedures were used to equate the raw composite scores on the alternate form to raw composite scores on the main form. Scale scores on the main form were developed by transforming the composite scores to be approximately normally distributed with a mean of 35, a

standard deviation of 10, and a range of 0 to 70. It is important to note that scale scores are not used operationally, but were included for research purposes.

For the main form, composite scores were transformed to AP grades of 1 to 5 using operational cut scores that were developed using a standard setting process. For the alternate form, composite scores were transformed to AP grades of 1 to 5 by equating raw composite scores on the alternate form to raw composite scores on the main form; these equated scores were then transformed to scale scores and AP grades using a conversion table.

Scale linking and equating were conducted using the released version of *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009), which is a set of open-source C functions. Item parameter and examinee ability estimates were obtained in MULTILOG, PARSCALE, flexMIRT, and IRTPRO for all four model combinations, using three different item prior settings, which resulted in a total of 12 estimated equating relationships per program.

Evaluation Criteria

Estimated equating relationships were analyzed for raw composite scores and unrounded scale scores through the use of difference plots. Classification consistency was computed for AP Grades by computing an overall exact agreement percentage (EAP) statistic for each condition. Overall EAPs were computed as the percentage of alternate form raw score points that converted to the same AP grade as the criterion method on the old form scale. Each score point on the alternate form was given equal weight in computing the overall EAPs, meaning that the agreement statistics were not weighted by the sample density.

Results from IRTPRO were used as the criterion to which results from the other three programs were compared. This decision was arbitrary and does not indicate that IRTPRO represents a gold standard among the studied programs. Within-program comparisons were made by using the baseline setting as a criterion. Therefore, within each program, and for each model combination, equating relationships resulting from the other two prior settings were compared to the relationship using the baseline setting. It is important to mention that there was no absolute or true criterion in the current study, because the focus was to compare differences among and within the four calibration programs.

Results

This section begins with an analysis of general item, test, and sample characteristics. Next, results pertaining to estimated IRT true score equating relationships for raw scores and

unrounded scale scores are presented. Last, overall EAPs for between-program comparisons (using IRTPRO as the criterion) and within-program comparisons (using the baseline setting as the criterion) are presented for AP grades.

Item Characteristics

When conducting IRT true score equating with models that incorporate the pseudo-guessing parameter, true scores can only be associated with a value of θ between scores of ∑ cj and K, where K represents the number of items on the test and the summation taken over all items. The lowest possible true score with the 3PL model is the sum of the *c*-parameter estimates. For both forms, observed scores lower than $\sum c_j$ were converted to possible true scores using an ad hoc procedure described by Kolen and Brennan (2014, p. 196). The sums of the *c*-parameter estimates for both forms are presented in Table 3.

For the 3PL + GR combination, the sum of the *c*-parameter estimates was around 18 on both forms, although the sum was consistently lower for the new form. It appears as though, on average, the estimated *c*-parameters from PARSCALE tended to be slightly lower than the other programs. The sums of the *c*-parameter estimates for the 3PL + GPC combination were slightly higher and were near 19 across the majority of programs and forms. However, the sums of the *c*-parameter estimates were generally lower by about one point for PARSCALE. In general, it appears that IRT true scores are undefined for observed scores of 18 and below, and equating results for scores in this range are found using the extrapolation procedure described earlier.

Sample Characteristics

Prior to examining the equating results, descriptive statistics for the main and alternate form samples were inspected and can be found in Table 4. By looking at the means of the common item scores, it appears that the alternate form examinees (*Mean* = 14.24) were slightly higher achieving than the main form examinees (*Mean* = 13.75). In general, the distributions for the common item scores were very similar across the main and alternate form samples.

Equating Relationships for Raw Composite Scores

In this section, results for estimated equating relationships for raw composite scores are presented. Results for between-program comparisons are presented first and are followed by comparisons made within each program (i.e., between different item prior settings).

Between-program comparisons. The estimated IRT true score equating relationships for raw scores (and scale scores) were exactly the same for flexMIRT and IRTPRO, and therefore

only results for IRTPRO are presented in this study. This was not surprising since there were no differences found in the estimated item parameters produced by IRTPRO and flexMIRT (see Peterson et al., 2014 in Chapter 4 for details). For each of the four model combinations, the estimated raw score equating relationships resulting from each calibration program are plotted in Figures 1 to 4, for the 2PL + GPC, 2PL + GR, 3PL + GPC, and 3PL + GR models, respectively.

Regardless of which model combination was used, the main form appeared easier than the alternate form for examinees with raw composite scores of approximately 45 and higher. When the 2PL model was used with the MC items, the estimated equating relationships were very similar across all software programs (Figures 1 and 2). Specifically, for the 2PL + GPC model, the equating relationships from all programs were basically identical, and for the 2PL + GR model, only small differences were seen between MULTILOG and the other programs.

When the 3PL model was used, discrepancies between programs were more noticeable and can be seen in Figures 3 and 4. Greater differences between programs were seen when the 3PL + GR model was used. For the 3PL + GR model, the estimated equating relationships from IRTPRO and MULTILOG were more similar to one another, whereas results from PARSCALE appeared quite different. The difference between MULTILOG and IRTPRO was found only at the low end of the ability scale and was quite small. The most noticeable difference, which was between IRTPRO and PARSCALE, was found at raw composite scores between 20 and 80. For the 3PL + GPC model, small differences between IRTPRO and PARSCALE were seen at the lower end of the ability distribution, and the raw score equating relationships from IRTPRO and MULTILOG were very similar.

Within-program comparisons. Differences in estimated equating relationships related to the use of different item prior settings are presented separately for IRTPRO, MULTILOG, and PARSCALE, and in that order. For each program, differences are broken down according to the four model combinations.

The comparisons of IRTPRO estimated equating relationships for raw scores using different item prior settings can be found in Figures 5 to 8, for the 2PL + GPC, 2PL + GR, 3PL + GPC, and 3PL + GR models, respectively. For each model combination, the use of different item priors did not have a noticeable effect on the shapes of the estimated equating relationships for raw scores, as they were indistinguishable from one another.

The comparisons of MULTILOG estimated raw score equating relationships under different prior settings are shown in Figures 9 to 12 for the 2PL + GPC, 2PL + GR, 3PL + GPC, and 3PL + GR models, respectively. For each model combination, the use of different item prior settings did not seem to affect the shapes of the estimated equating relationships for raw scores, as they overlapped each other for the majority of the raw score scale.

The comparisons of PARSCALE estimated raw score equating relationships resulting from the use of different prior settings can be seen in Figures 13 to 16 for the 2PL + GPC, 2PL + GR, 3PL + GPC, and 3PL + GR models, respectively. The use of different item prior settings led to basically the same estimated equating relationships for all model combinations except for the 3PL + GPC. For the 3PL + GPC model, the estimated equating relationship from using the baseline setting was slightly different than the ones estimated by using the alternate two settings. Differences between the baseline setting and the other two settings were found for nearly all score points (with the exception of alternate form scores > 130); however differences were never very large.

Equating Relationships for Unrounded Scale Scores

Similar to raw scores, only the estimated scale score equating relationships for IRTPRO, MULTILOG, and PARSCALE are presented in this study. Results for flexMIRT were not included because they were identical to IRTPRO. In the difference plots for equated scale scores, the horizontal axis still represents the alternate form raw composite score, however the y-axis now represents the difference in equated unrounded scale scores between IRTPRO and the other two programs. Specifically, the difference was calculated as the equated unrounded scale scores from MULTILOG/PARSCALE minus the equated unrounded scale score from IRTPRO.

Between-program comparisons. The differences in estimated scale score equating relationships between software programs can be found in Figures 17 to 20, for the 2PL + GPC, 2PL + GR, 3PL + GPC, and 3PL + GR models, respectively.

In general, greater similarity was seen among the estimated equating relationships for unrounded scale scores than for raw scores. For the 2PL + GPC and 2PL + GR models, the use of different programs did not lead to any noticeable difference in the scale score equating relationships (Figures 17 and 18). When the 3PL + GPC and 3PL + GR models were used, small differences were found at the lower and higher ends of the ability distribution, which is also consistent with the findings for raw scores (Figures 19 and 20).

Within-program comparisons. Estimated scale score equating relationships related to different item prior settings are presented in Figures 21 to 32, separately for IRTPRO, MULTILOG, and PARSCALE. Within each program, comparisons are made for each of the four model combinations. The difference in equated unrounded scale scores were plotted by subtracting results from the baseline setting from results using the alternate settings.

For IRTPRO, regardless of the IRT model combination, different item prior settings appeared to produce almost identical equating relationships for unrounded scale scores, which is consist with findings from the estimated raw score equating for this program.

For MULTILOG, when the 2PL models were used for MC items, differences between the baseline setting and the alternate two settings were not visually noticeable. For the 3PL + GPC model combination, small difference was observed between the baseline setting and *c*-prior only setting only at the higher end of the ability distribution. For the 3PL + GR models, results between settings were slightly different at the lower end of the ability distribution; however, this might not be important because equated scores lower than the sum of *c*-parameter estimates were determined by linear interpolation

For PARSCALE, when the 2PL models were used for the MC items, the use of different item prior settings did not seem to affect the equating relationships for scale scores, however some differences were found when the 3PL models were used. For the 3PL + GPC model, differences between the baseline setting and the alternate two settings were seen along the majority of score scale, reaching a maximum difference of 0.5 scale score points around an alternate form raw scores of 18 (sum of *c*-parameters) and scores slightly lower than 150 (i.e., near a perfect score). These differences were also consistent with the raw score differences. For the 3PL + GR model, even though no significant differences were found for raw scores, small differences were found in the scale score relationships at the lowest and highest ends of the ability distribution. This inconsistency between raw and scale score equating relationships might not be surprising, although, up to this point, the tendency has been for raw score results to carry over to scale scores. One possible explanation for the difference seen with PARSCALE results is that a nonlinear transformation was used to convert raw composite scores to scale scores.

Agreement Statistics for AP Grades

Because AP grades (and not raw scores or scale scores) are ultimately used by institutions of higher education to make decisions, it was important to evaluate any differences in AP grades

that resulted from the use of different software programs and/or different item prior settings. Therefore, agreement statistics were computed to determine whether the differences observed between estimated IRT true score equating relationships carried practical implications for examinees. To answer this question, overall exact agreement percentages (EAPs) were computed between equated AP grades using a criterion method and equated AP grades associated with each remaining condition. For between-program comparisons, the equated AP grades from IRTPRO served as the criterion, whereas for within-program comparisons the equated AP grades from the baseline setting served as the criterion.

Between-program comparisons. For all model combinations, the overall EAP between IRTPRO and the other software programs can be found in Table 5. Perfect exact agreement between IRTPRO and the other two programs was found for all four model combinations. Therefore, the small differences in raw score equating relationships did not carry over to AP Grades. Put another way, in this study, the choice to use either IRTPRO, MULTILOG, or PARSCALE had no practical implications for equated AP Grades.

Within-program comparisons. Within each program, the overall EAPs for AP grades were found by comparing the two alternate item prior settings to the baseline prior setting and results can be seen in Tables 6 and 7. In Table 6, EAPs between the baseline setting and the "*a*- and *c*-prior" setting are displayed for each program. Perfect agreement was attained for all model combinations with IRTPRO and MULTILOG. When PARSCALE was used, all conditions except for the 3PL + GPC model combination resulted in perfect agreement. Still, the overall EAP when for the 3PL + GPC model combination was 99.34%, and implies that only 1 of the 151 raw score points was inconsistent as a result of using different item prior settings.

In Table 7, overall EAPs between the baseline setting and the "c-prior only" setting are displayed for each program. Again, perfect agreement was found for IRTPRO and MULTILOG, meaning that each raw score on the alternate form equated to the same AP Grade on the main form scale, for all model combinations. The equated AP Grades from PARSCALE were also very consistent and again, the only discrepancy was found for the 3PL + GPC model combination which had an overall EAP of 99.34%.

In general, at the AP grade level, there was a high degree of consistency between the equatings resulting from the baseline and alternate calibration settings. The overall percent agreement, collapsing across the five AP grade levels, ranged from 99.34 to 100 percent. The

greatest consistency was seen with IRTPRO and MULTILOG, but was followed very closely by PARSCALE.

Discussion

Comparisons across four different IRT software programs were made using MULTILOG, PARSCALE, IRTPRO, and flexMIRT. Comparisons were typically made with reference to results produced from IRTPRO since it is a newer program and was developed with the intent to replace MULTILOG. In general, parameter estimates and equating results obtained from flexMIRT and IRTPRO were identical for all IRT model combinations considered. This was not surprising since the two programs have overlapping theoretical frameworks and authors. Therefore, results from flexMIRT were excluded from the current study because they did not contribute any unique information. However, small differences were found between flexMIRT/IRTPRO and the other two software programs.

Effect of Using Different Software Programs

In general, estimated equating relationships for raw and scale scores from IRTPRO and MULTILOG were very similar, and appeared nearly identical for the majority of model combinations. The only difference observed between these two programs was when the 3PL + GR model combination was used, and even then, differences were limited to the extreme ends of the ability distribution and were quite small. This finding supports the use of IRTPRO as MULTILOG's successor and suggests that few differences can be expected when transitioning between programs. Because results from IRTPRO and flexMIRT were basically identical, this finding also suggests that users could adopt flexMIRT as an alternative to IRTPRO without expecting additional complications.

Differences in estimated raw and scale score equating relationships were found between IRTPRO and PARSCALE only when the 3PL model was used. The greatest difference was found for the 3PL + GR model combination. For this model combination, the equated scale scores from PARSCALE were higher than those from IRTPRO for alternate form raw scores of 90 and below, and were lower than those from IRTPRO for raw scores of 110 and above. This trend was also seen for the comparison of raw score equating relationships. When the 3PL + GPC model combination was used, the differences between IRTPRO and PARSCALE were more subtle and were limited to the lower end (i.e., between raw scores of 18 and 50) of the

ability distribution. The maximum differences between scale scores for examinees with raw scores in this range was approximately 0.5 scale score points.

In order to determine whether any of the aforementioned differences between software programs carried practical significance for examinees, overall EAPs were computed for rounded AP Grades. It turned out that differences did not carry over into AP Grades and therefore, examinees would have received the same AP Grade regardless of which software program was used. This is an important finding for operational testing programs that group student performance into categories such as Advanced, Proficient, Basic, and Below Basic. Similar to the AP program, these reporting scales are much broader and, as a result, scores of this type (in comparison to more precise scales) are less likely to be affected by minor changes to calibration settings or by adopting a different software program.

Effect of Using Different Prior Settings

Three sets of item prior settings were used with MULTILOG, PARSCALE, and IRTPRO. Effects of using different item priors were investigated within each program with regard to equating relationships for raw and unrounded scale scores, along with agreement statistics.

In general, the use of different item priors did not appear to affect the estimated IRT true score equating relationships to a great extent. However some small differences were observed. IRTPRO was the only program that appeared completely unaffected by the use of different item priors. In other words, the estimated equating relationships for raw and unrounded scale scores were visually indistinguishable across the three item prior settings.

Similarly, there were no apparent differences across item prior settings for MULTILOG and PARSCALE when the 2PL model was used. For MULITLOG, and with the 3PL + GPC model combination, the raw and scale score equating relationships associated with the baseline and *c*-prior only settings were slightly different for scores near the upper end of the distribution. When the 3PL + GR model was specified in MULTILOG, a slight difference in estimated equating relationships between the baseline versus the alternate two prior settings was found. This difference was small and constrained to raw scores between approximately 15 and 40.

A similar pattern was observed with PARSCALE such that when the 3PL model was used, small differences emerged. However, the pattern was consistent such that the equating relationships obtained by using the alternate prior settings were very similar to one another, but

slightly different than those found by using the baseline setting. For the 3PL + GPC model, the differences between the baseline and alternate prior setting raw score equating relationships appeared small and consistent across the majority of the ability distribution. However, the differences in the scale score equating relationships appeared larger near the lower and upper ends of the ability distribution and less different near the middle. When the 3PL + GR model was used, the raw and scale score equating relationships compared across item prior setting appeared very similar for the most part, although a very small difference was observed near raw scores of 18 (i.e., sum of *c*-parameters) and 140. Again, the baseline setting produced a slightly different equating relationship from the alternate settings, and the alternate settings were virtually the same.

The use of different item prior settings did not have a large impact on the consistency of AP Grade assignments. Actually, for IRTPRO and MULTILOG, the use of different item priors led to the exact same AP Grades being assigned for all model combinations. The same was true for PARSCALE for all model combinations except for the 3PL + GPC. For this model combination, 1 of the 151 raw score points resulted in a different AP Grade depending on which item prior setting was used. This resulted in an overall EAP of 99.34% between the baseline setting and the alternate settings. It is important to mention that the agreement statistics used in the current study were unweighted and therefore each attainable raw score value on the alternate form, received equal weight. Computing agreement statistics in this manner could disguise the significance of the discrepancies, and is worth future consideration.

This study has demonstrated that by changing item prior settings there is a potential for a discrepancy to arise in equated scores, even when the scale is quite broad such as with the AP Grade scale. This finding is likely of more concern to testing programs that use IRT equating methods and that have small examinee populations because the use of item priors is known to influence estimation more with small samples. However, the current conclusions should be interpreted with caution since they are based on a single exam and may not be generalizable to all mixed-format exams. Future studies should seek to determine whether the results found here can be generalized to other subject areas and whether any differences observed hold practical significance when weighted agreement statistics are used.

Conclusions

Given the number of IRT models and IRT software programs available today, it is important to empirically compare psychometric differences that could result from the decision to choose one over another. This type of comparison was especially important with the release of IRTPRO and flexMIRT and the subsequent retirement of MULTILOG. A notable finding in this study was the high degree of consistency in AP Grade assignments across calibration programs and item prior settings. In general, many fewer differences were found between estimated equating relationships when the 2PL model was used in comparison to the 3PL model, and may be related to the fewer number of parameters being estimated. Comparisons between programs revealed a slightly higher consistency between IRTPRO and MULTILOG than between IRTPRO and PARSCALE. This finding supports the replacement of MULTILOG with IRTPRO, and given the overlap between IRTPRO and flexMIRT results, suggests that flexMIRT is an equally suitable replacement for MULTILOG.

The current study was limited by the fact that the "true" equating relationship was unknown since real data were used. Therefore, it was impossible to determine which software program or item prior settings resulted in the most accurate equating relationships. For this reason, it may be desirable to conduct a similar study when the true equating relationship and true IRT models are known. Last, since this study included data from only the AP Biology exams, future replication studies should be conducted using different subject areas.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

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.

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

Cai, L., Thissen, D. J., & du Toit, S. (2011). *IRTPRO* (Version 2.1) [Computer program]. Mooresville, IN: Scientific Software.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E., & Bock, R. D. (2003). *PARSCALE* (Version 4.1) [Computer program]. Mooresville, IN: Scientific Software.

Peterson, J., Zhang, M., Pak, S., Wang, S., Wang, W., Lee, W., & Kolen, M. J. (2014). A comparison of several item response theory software programs for calibrating mixed-format exams. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3).* (*CASMA Monograph No. 2.3*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Thissen, D. J., Chen, W.-H., & Bock, R. D. (2003). *MULTILOG* (Version 7.0) [Computer program]. Mooresville, IN: Scientific Software.

Table 1

Prior Distributions Used During Item Parameter Estimation

Condition	Slope (*a*)	Location (*b*)	Pseudo-guessing (*c*)
Baseline	Normal (1.133, 0.604)	None	Logit Normal (-1.39, 0.5)
a- and *c*-prior	Normal (1, 1)	None	Logit Normal (-1.39, 0.5)
c-prior only	None	None	Logit Normal (-1.39, 0.5)

Note. PARSCALE used *Beta* (5, 20) on the *c*-prior for all conditions.

Table 2

General Calibration Settings Used During Item Parameter Estimation

Setting	MULTILOG	PARSCALE	IRTPRO	flexMIRT
# of cycles for E steps	3,000	3,000	3,000	3,000
# of cycles for M steps	3,000	3,000	3,000	3,000
Convergence criterion for E steps	0.0001	0.0001	0.0001	0.0001
Convergence criterion for M steps	0.0001	0.0001	0.0001	0.0001
# of quadrature points	49	49	49	49
Quadrature Range	[-6, 6]	[-6, 6]	[-6, 6]	[-6, 6]

Table 3

Sum of c-Parameter Estimates Using the Baseline Item Prior Setting

Model & Form	MULTILOG	PARSCALE	IRTPRO	flexMIRT
3PL + GR				
Main	18.25	17.68	18.25	18.25
Alternate	17.91	17.11	17.73	17.73
3PL + GPC				
Main	18.61	18.25	18.61	18.61
Alternate	18.26	17.33	18.26	18.25

Table 4

Descriptive Statistics for Equating Samples ($N_{Main} = N_{Alt.} = 3{,}000$)

Source	Mean	Std. Dev.	Skewness	Kurtosis
Weighted Composite				
Main	80.31	30.96	-0.08	-0.92
Alternate	80.02	29.52	-0.15	-0.93
Common Items				
Main	13.75	4.76	-0.03	-0.71
Alternate	14.24	4.91	-0.15	-0.79

Table 5

Overall Exact Agreement Percentages for AP Grades Between IRTPRO and Other Programs

	Calibration Program	
Models	MULTILOG	PARSCALE
2PL + GPC	100	100
2PL + GR	100	100
3PL + GPC	100	100
3PL + GR	100	100

Note. Results were calculated using the Baseline item prior setting.

Table 6

Exact Agreement Percentages for AP Grades Between Baseline and "a- and c-prior" Settings

	Calibration Program		
Model	MULTILOG	PARSCALE	IRTPRO
2PL + GPC	100	100	100
2PL + GR	100	100	100
3PL + GPC	100	99.34	100
3PL + GR	100	100	100

Note. Priors used in the "*a*- and *c*-prior" setting were $a \sim$ Normal (1, 1) and $c \sim$ logit Normal (-1.39, 0.5) for MULTILOG and IRTPRO and $a \sim$ Normal (1, 1) and $c \sim$ Beta (5, 20) for PARSCALE.

Table 7

Exact Agreement Percentages for AP Grades Between Baseline and "c-prior only" Settings

	Calibration Program		
Model	MULTILOG	PARSCALE	IRTPRO
2PL + GPC	100	100	100
2PL + GR	100	100	100
3PL + GPC	100	99.34	100
3PL + GR	100	100	100

Note. Priors used in "*c*-prior only" setting were c ~ logit normal (-1.39, 0.5) for MULTILOG and IRTPRO and *c* ~ Beta (5, 20) for PARSCALE.

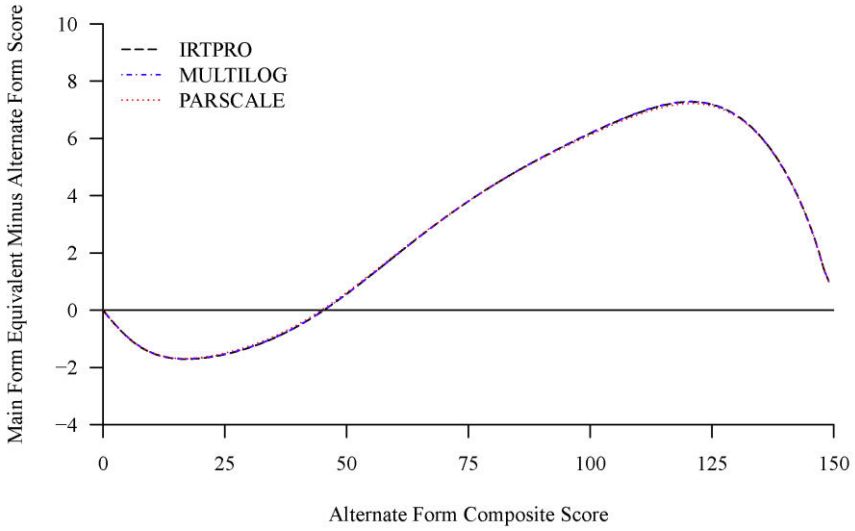


Figure 1. Between-program comparisons of raw score equating relationships for the 2PL + GPC model combination.

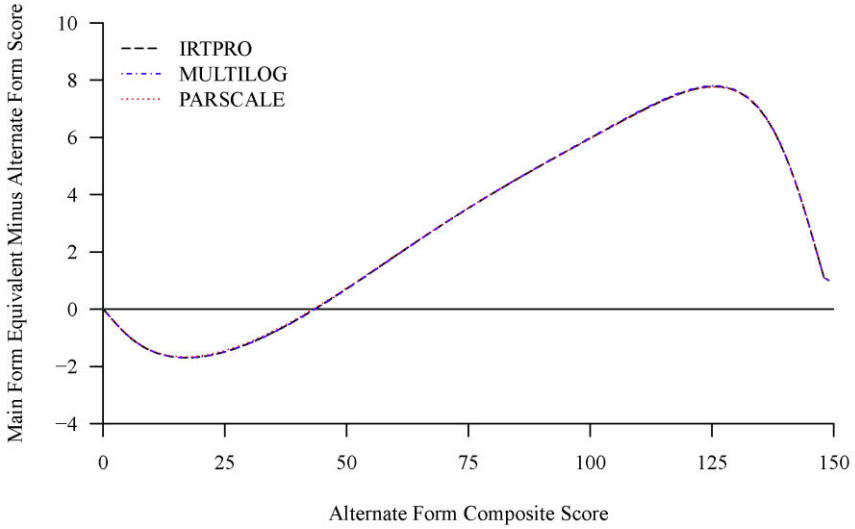


Figure 2. Between-program comparisons of raw score equating relationships for the 2PL + GR model combination.

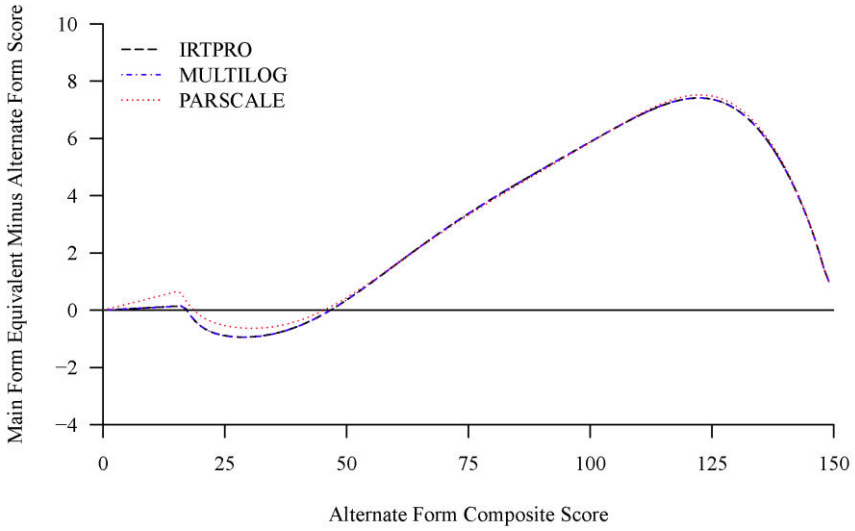


Figure 3. Between-program comparisons of raw score equating relationships for the 3PL + GPC model combination.

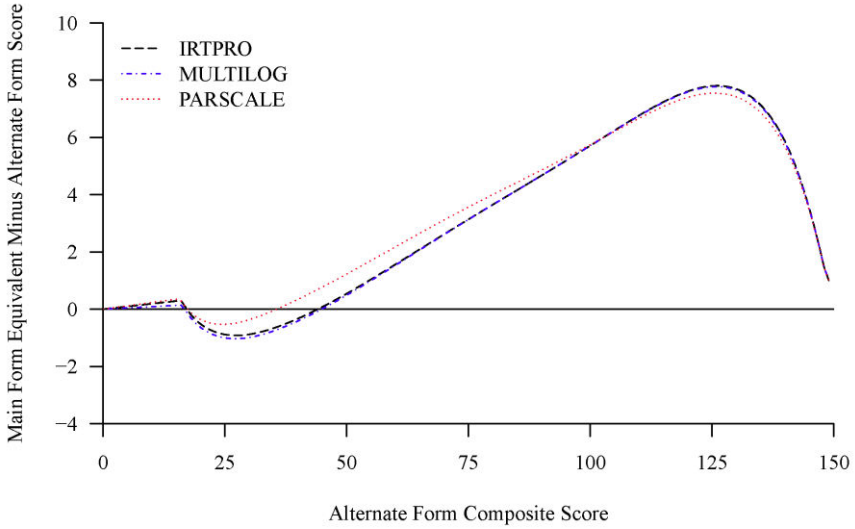


Figure 4. Between-program comparisons of raw score equating relationships for the 3PL + GR model combination.

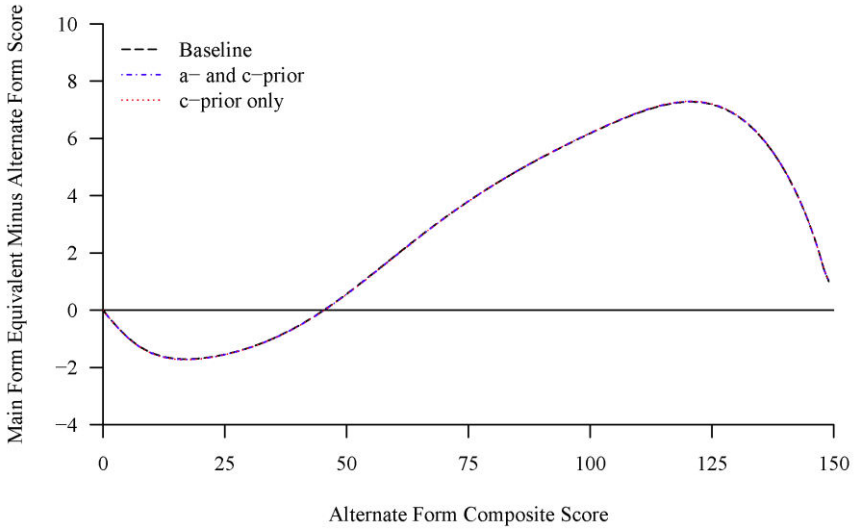


Figure 5. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GPC model combination in IRTPRO.

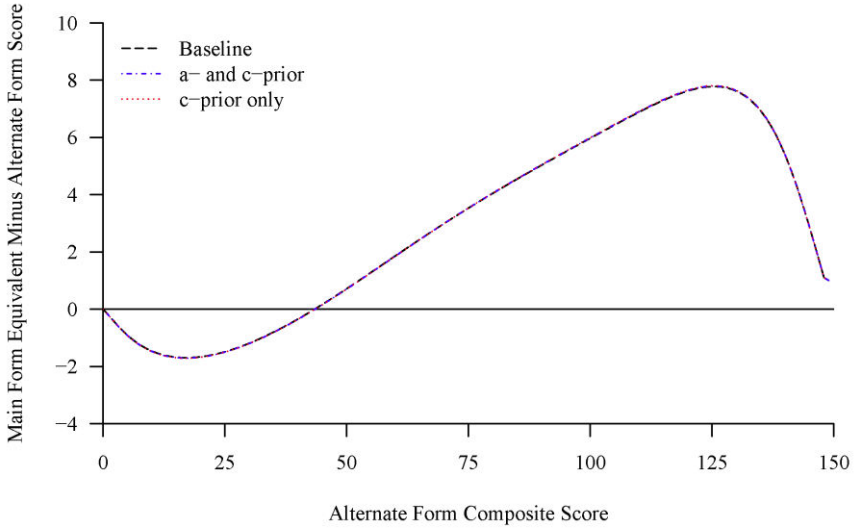


Figure 6. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GR model combination in IRTPRO.

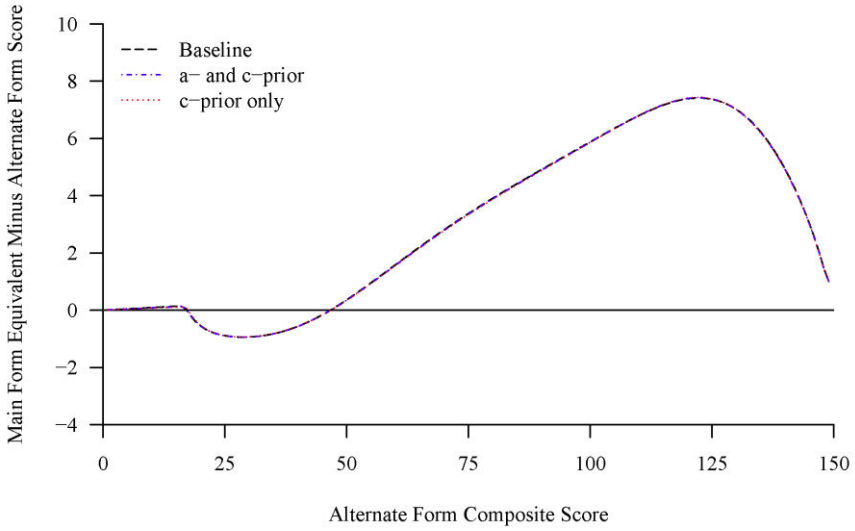


Figure 7. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GPC model combination in IRTPRO.

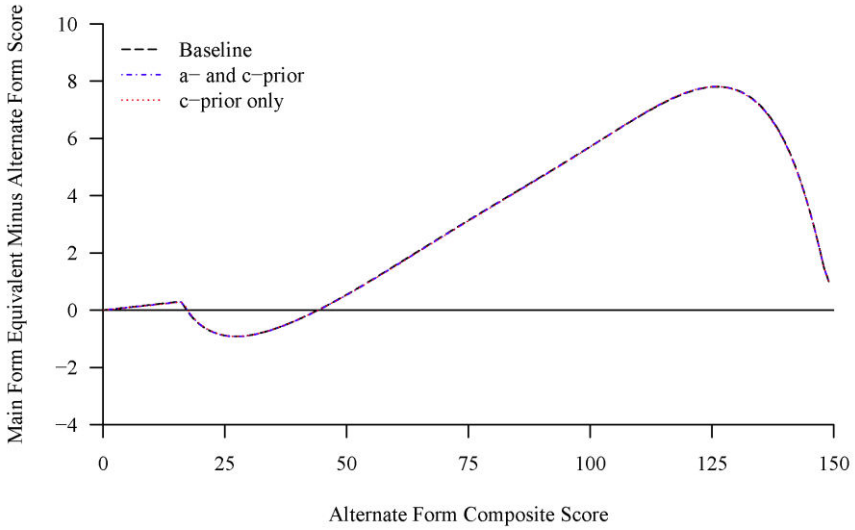


Figure 8. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GR model combination in IRTPRO.

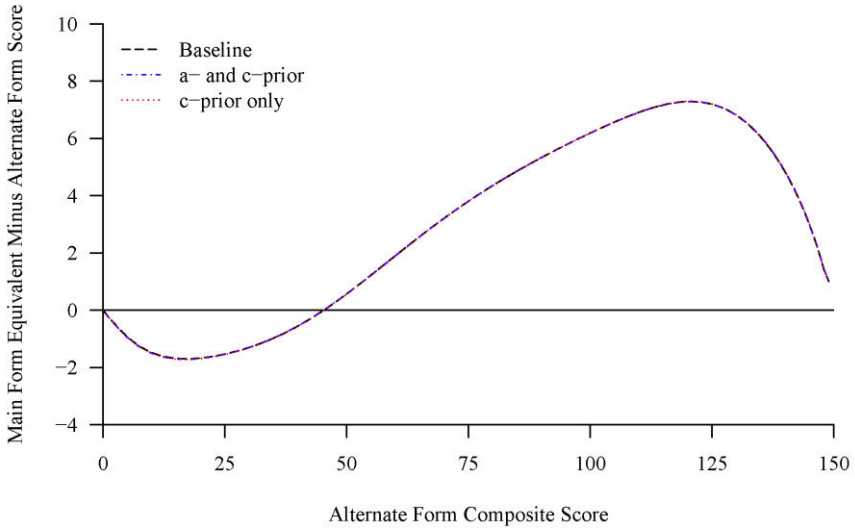


Figure 9. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GPC model combination in MULTILOG.

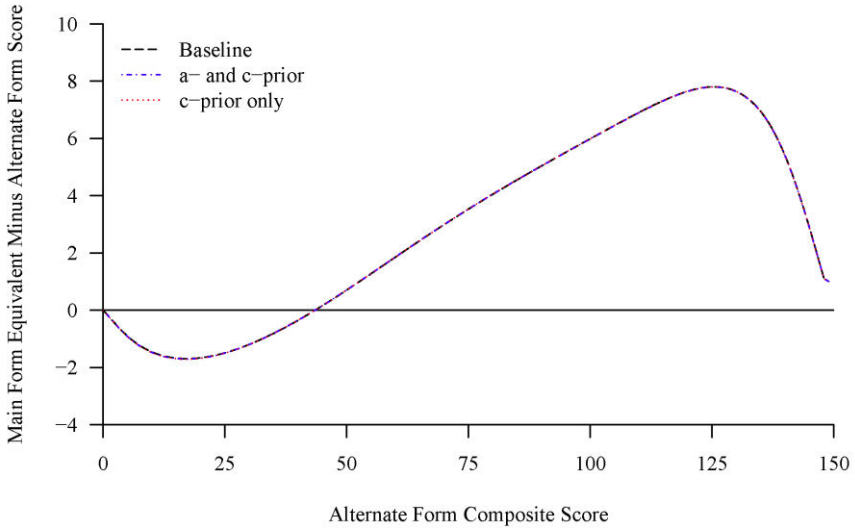


Figure 10. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GR model combination in MULTILOG.

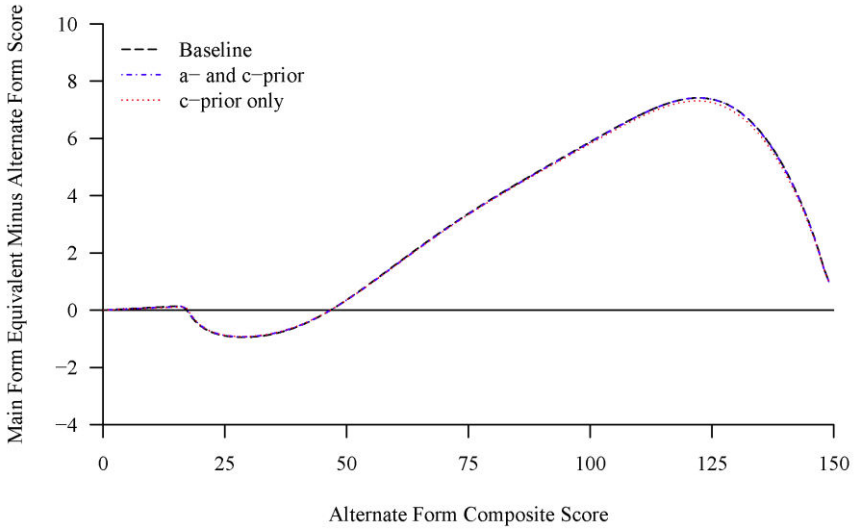


Figure 11. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GPC model combination in MULTILOG.

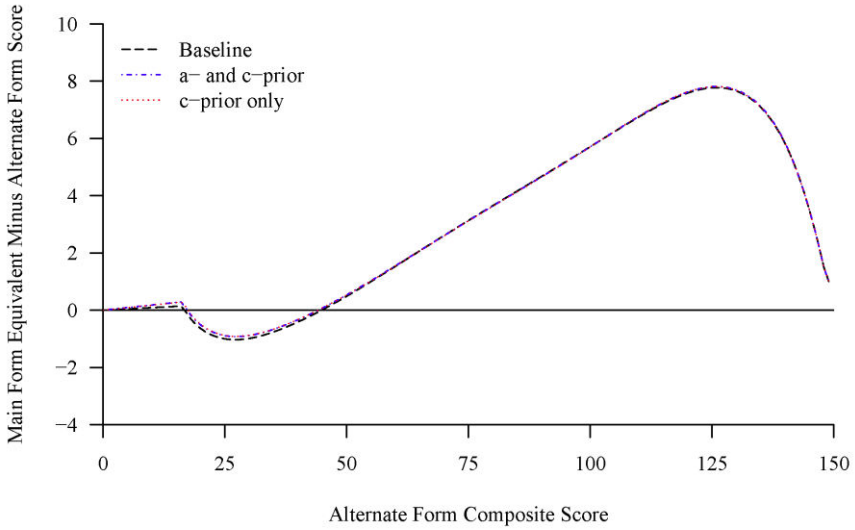


Figure 12. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GR model combination in MULTILOG.

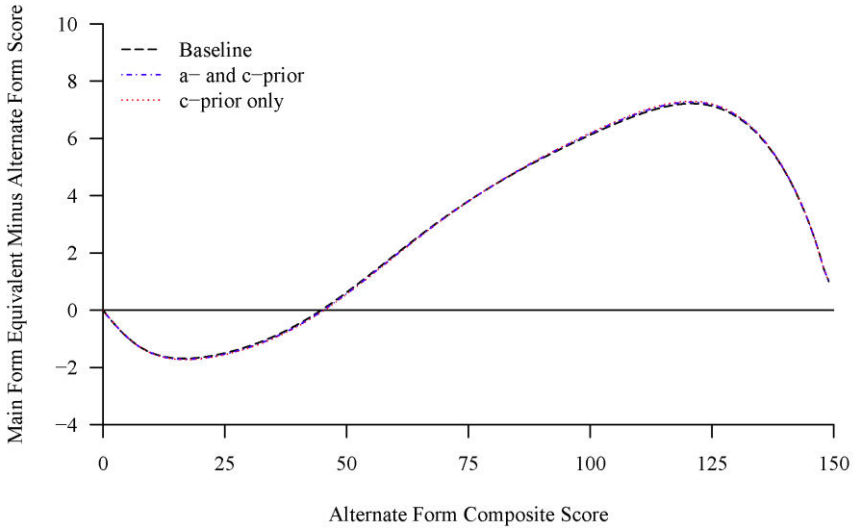


Figure 13. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GPC model combination in PARSCALE.

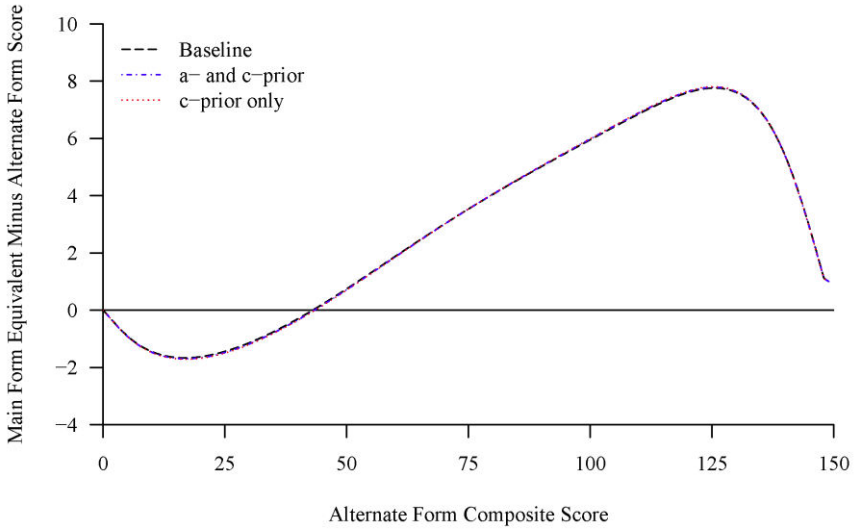


Figure 14. Comparisons of raw score equating relationships using each item prior setting for the 2PL + GR model combination in PARSCALE.

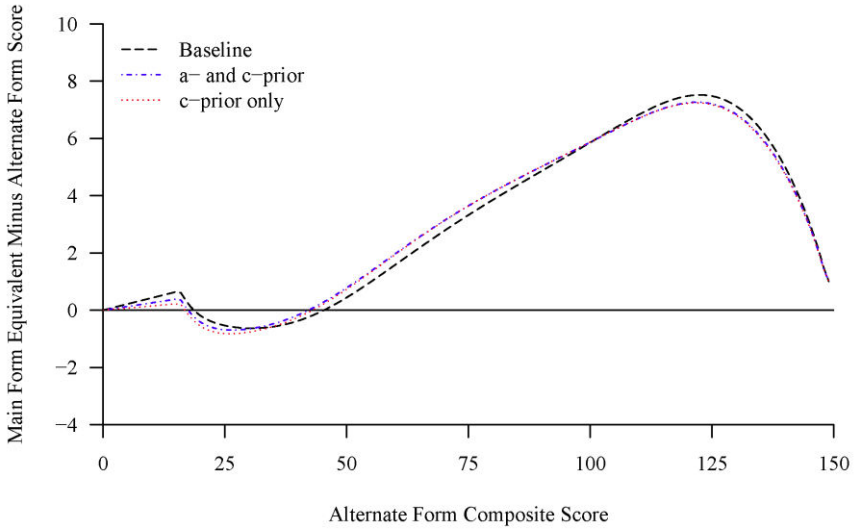


Figure 15. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GPC model combination in PARSCALE.

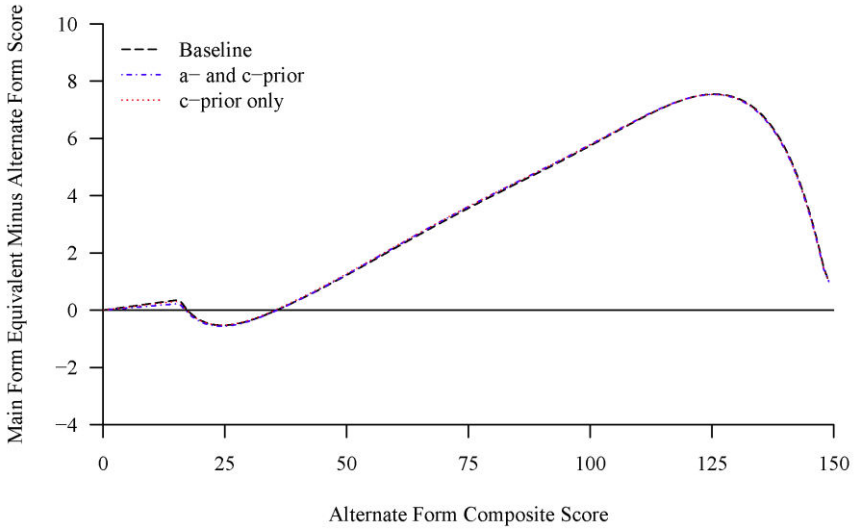


Figure 16. Comparisons of raw score equating relationships using each item prior setting for the 3PL + GR model combination in PARSCALE.

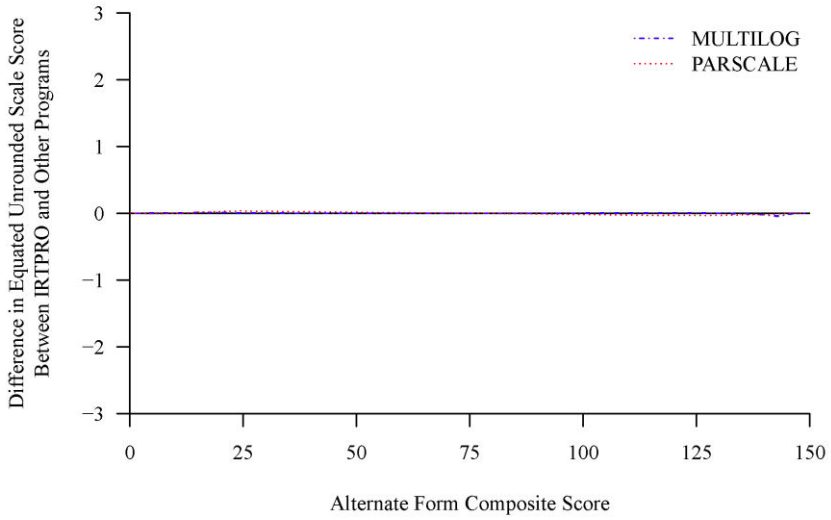


Figure 17. Between-program comparisons of scale score equating relationships for the 2PL + GPC model combination.

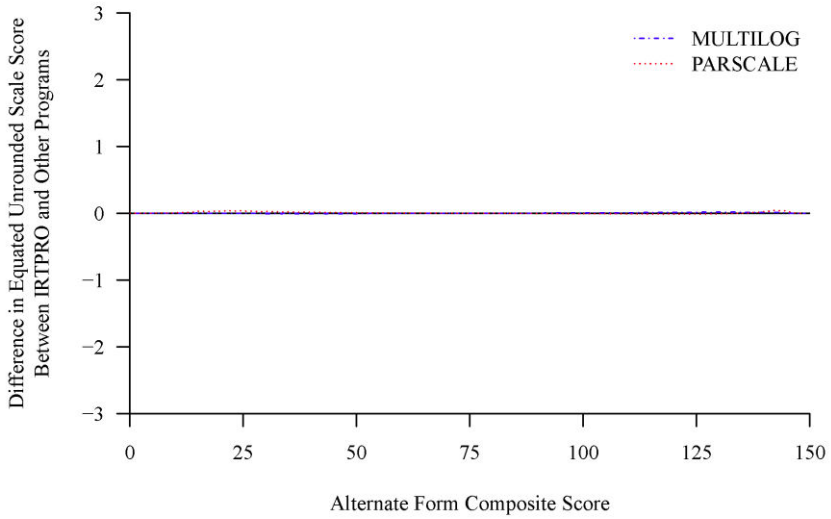


Figure 18. Between-program comparisons of scale score equating relationships for the 2PL + GR model combination.

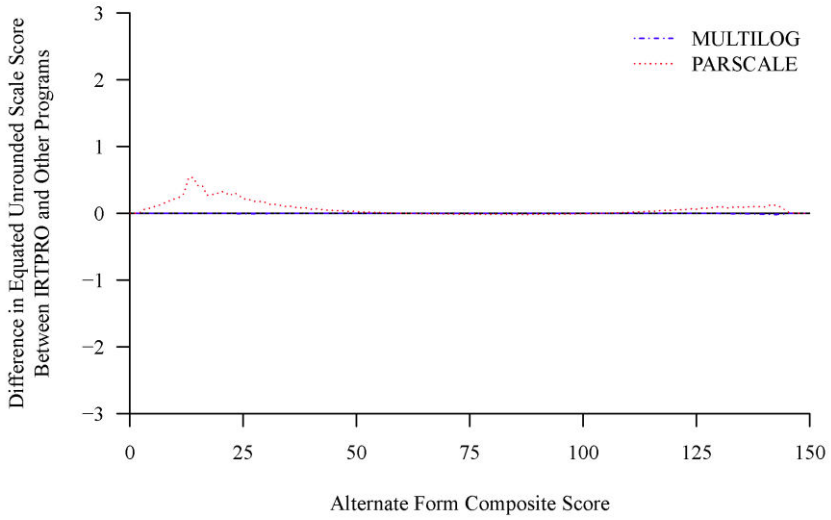


Figure 19. Between-program comparisons of scale score equating relationships for the 3PL + GPC model combination.

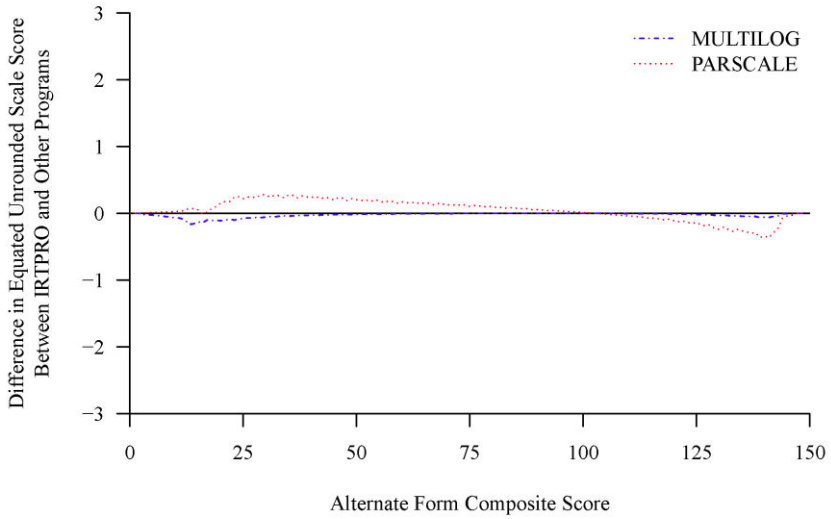


Figure 20. Between-program comparisons of scale score equating relationships for the 3PL + GR model combination.

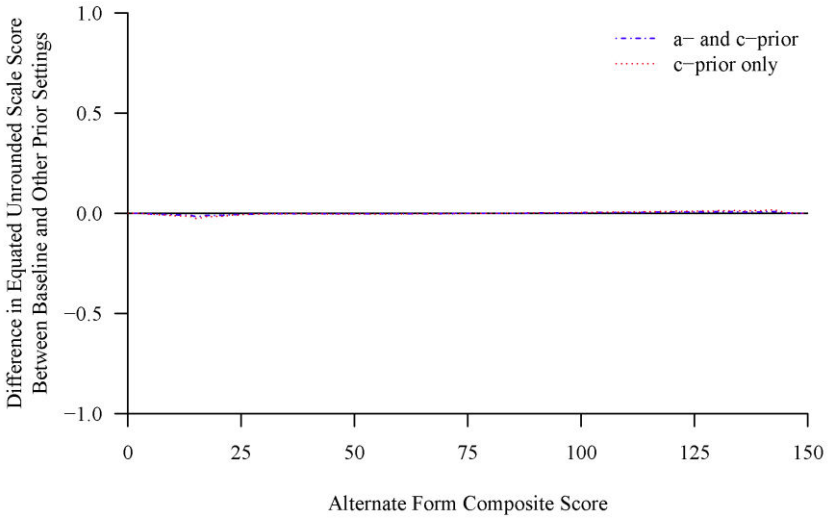


Figure 21. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GPC model combination in IRTPRO.

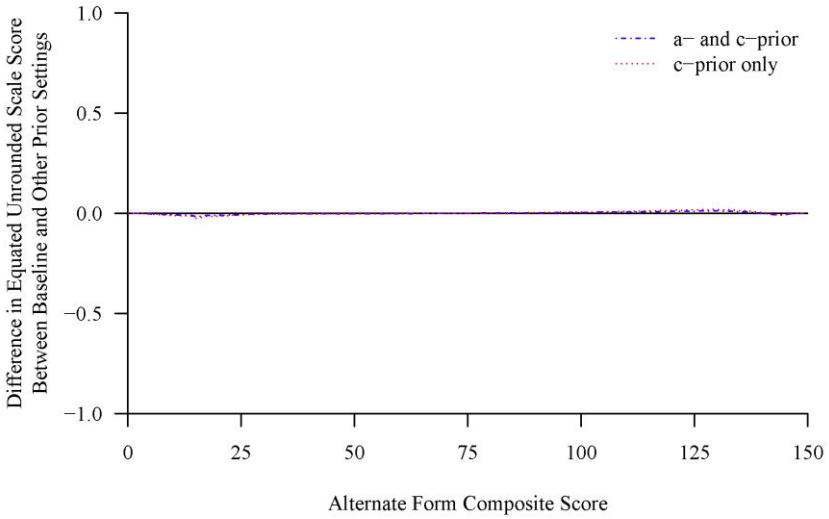


Figure 22. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GR model combination in IRTPRO.

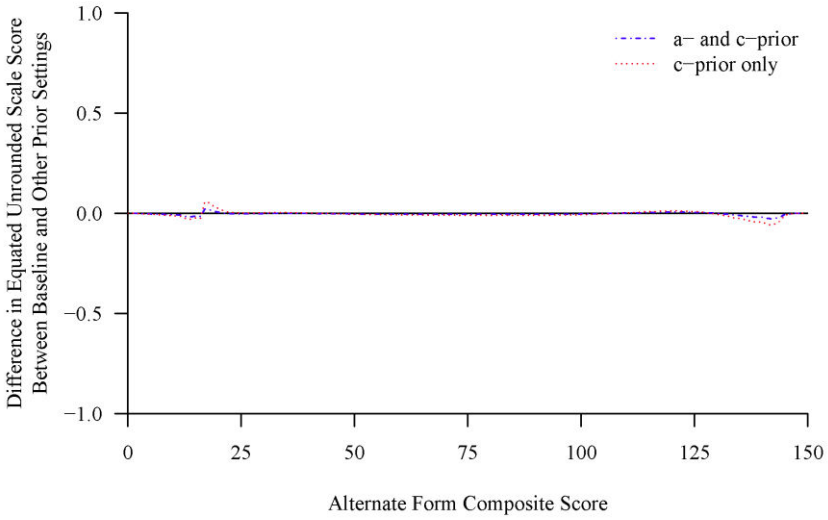


Figure 23. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GPC model combination in IRTPRO.

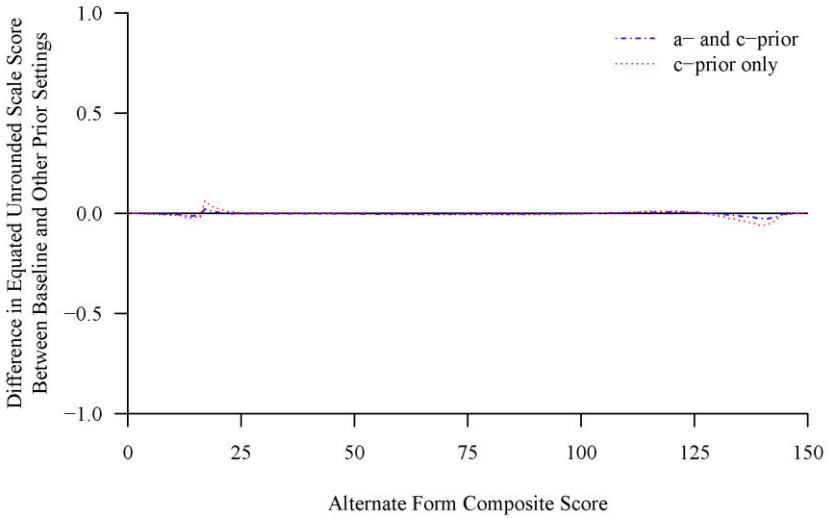


Figure 24. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GR model combination in IRTPRO.

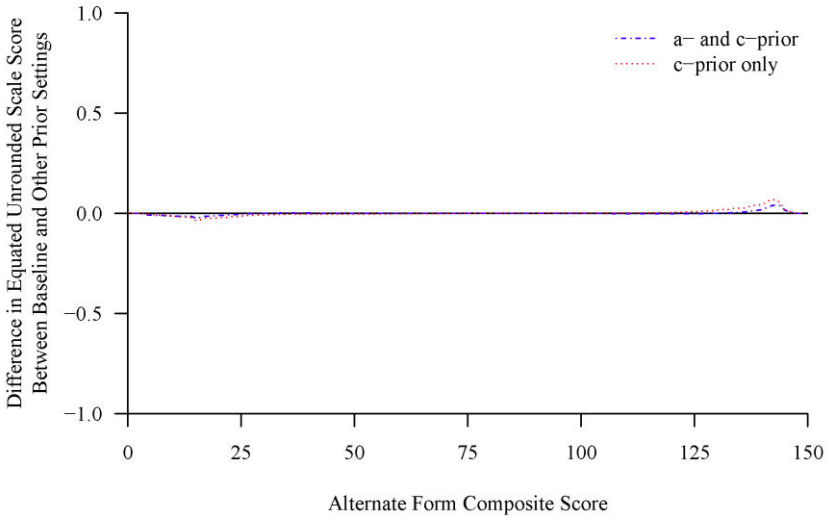


Figure 25. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GPC model combination in MULTILOG.

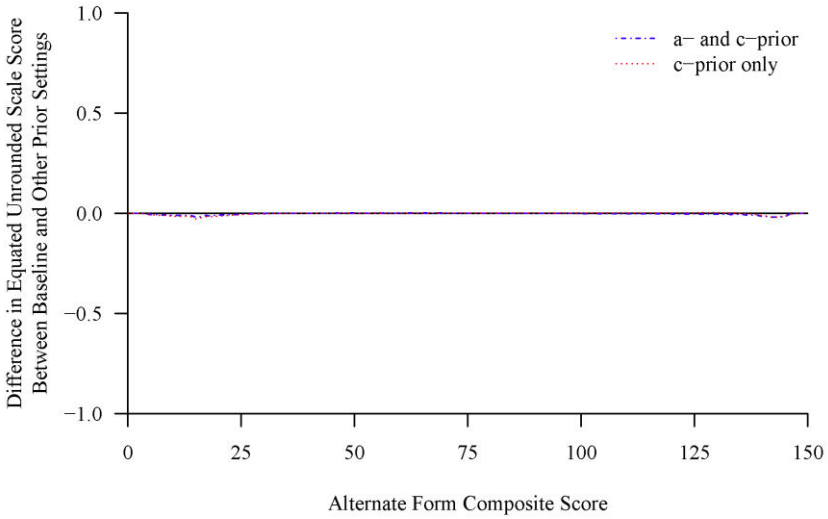


Figure 26. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GR model combination in MULTILOG.

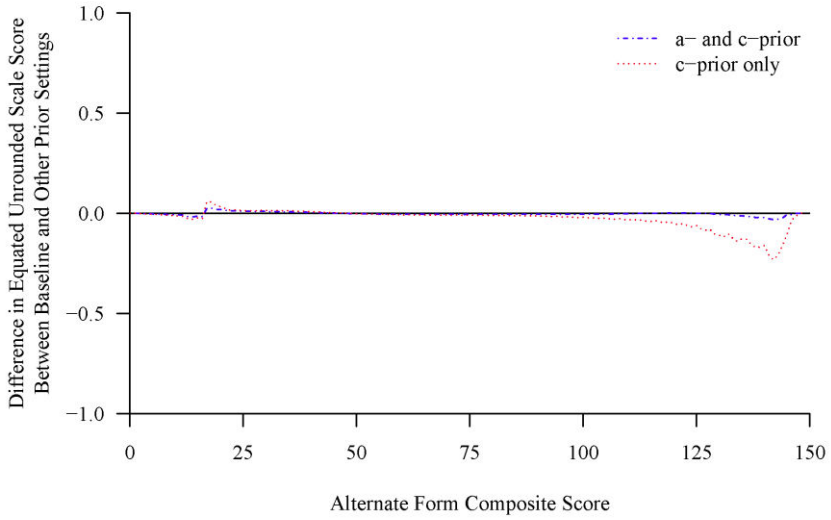


Figure 27. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GPC model combination in MULTILOG.

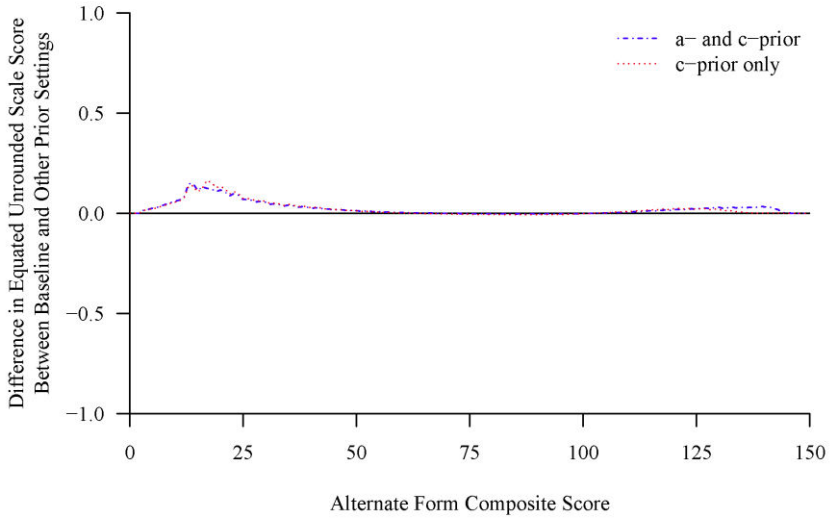


Figure 28. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GR model combination in MULTILOG.

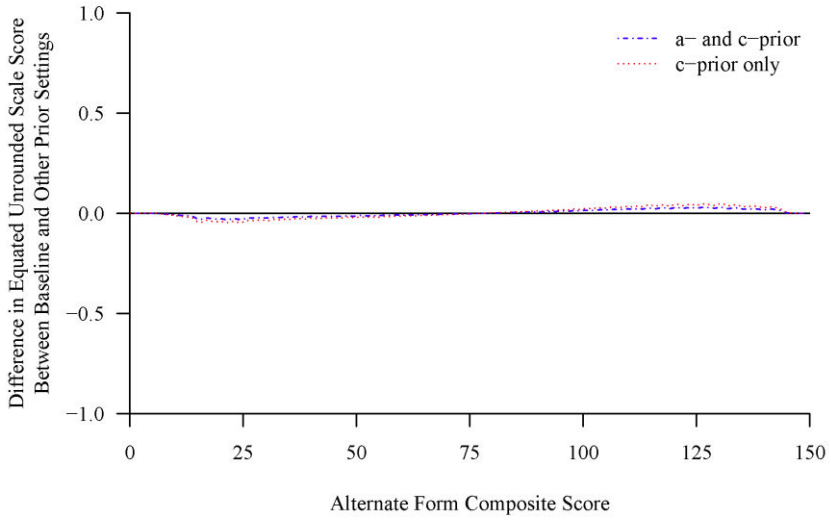


Figure 29. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GPC model combination in PARSCALE.

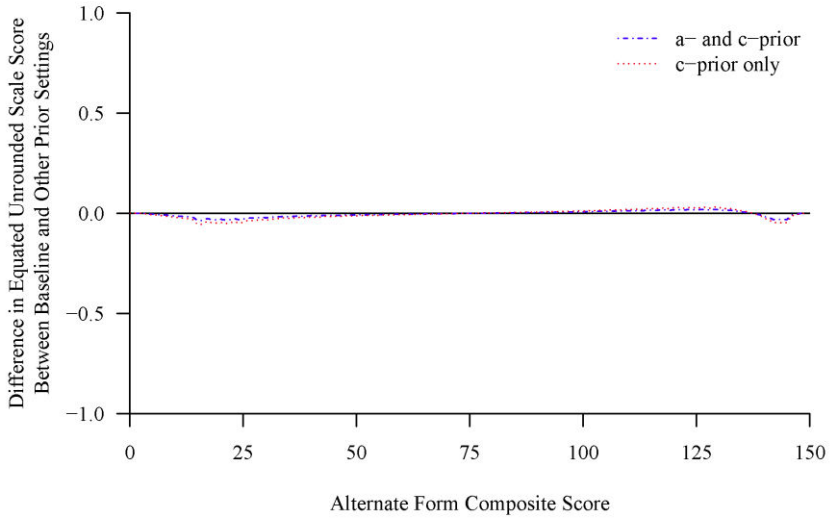


Figure 30. Comparisons of scale score equating relationships using each item prior setting for the 2PL + GR model combination in PARSCALE.

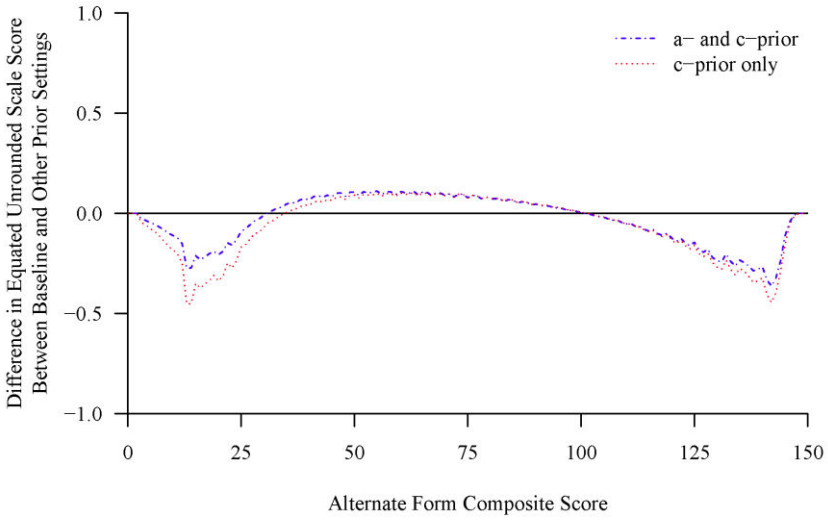


Figure 31. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GPC model combination in PARSCALE.

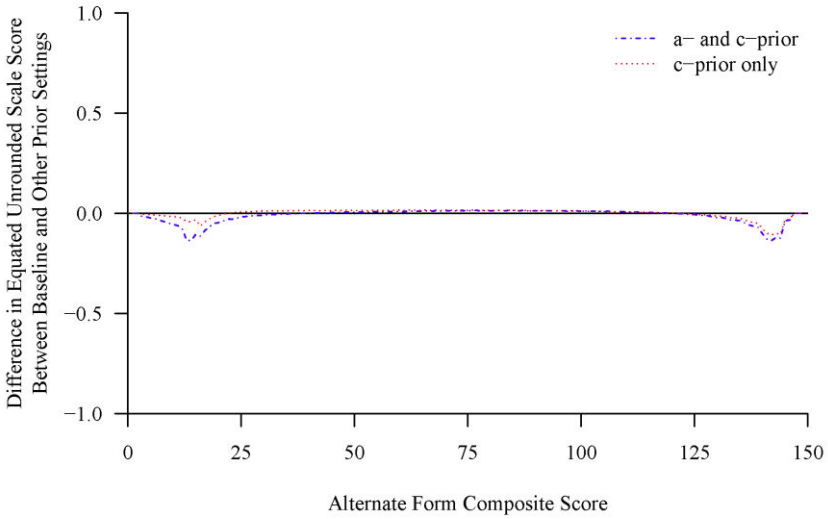


Figure 32. Comparisons of scale score equating relationships using each item prior setting for the 3PL + GR model combination in PARSCALE.

Chapter 6: A Comparison of Test Dimensionality Assessment Approaches for Mixed-Format Tests

Mengyao Zhang, Michael J. Kolen, and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

The use of mixed-format tests could result in a more complex dimensional structure compared to multiple choice-only tests and poses methodological challenges to dimensionality assessment. This study empirically compared a set of dimensionality assessment methods associated with exploratory factor analysis using data from mixed-format tests. Specifically, the number of dimensions was determined by four methods using principal components analysis. Dimensional structure was explored by item-level factor analysis considering a series of model-selection criteria. Tests from four different subject areas (English Language, Spanish Language, Comparative Government and Politics, and Chemistry) were used as illustrative examples. Results of this study confirm some general patterns of similarities and dissimilarities among results from different methods that were observed in previous studies. When the purpose of a dimensionality assessment was solely to check whether essential unidimensionality holds, some easy-to-implement eigenvalue rules provided almost identical results to those of more complicated factor analysis methods, but the latter could further help practitioners specify a test's dimensional structure. Both minimum average partial (MAP; Velicer, 1976) and parallel analysis (PA; Horn, 1965) methods appeared to be sensitive to the occurrence of trivial dimensions. Graphical approaches, such as scree plot and PA plot, could be effective when combined with other analytic methods. In addition, this study discussed some practical challenges in examining the test dimensionality, which could be helpful for practitioners when choosing between different dimensionality assessment methods.

A Comparison of Test Dimensionality Assessment Approaches for Mixed-Format Tests

In recent years, many large-scale testing programs have been using mixed-format tests that contain both multiple-choice (MC) and free-response (FR) items in order to combine the strengths of different item formats. However, a mixture of MC and FR items might result in a more complex dimensional structure compared to MC-only tests. For example, varying item formats might introduce an extra dimension unexpected to the test developers, as some previous studies found that MC and FR items tend to measure different types of cognitive skills (Cao, 2008). The growing number of mixed-format tests also poses methodological challenges to dimensionality assessment. Some widely used procedures for MC-only tests are not capable of handing dichotomous and polytomous data simultaneously, and therefore are unable to analyze response data for mixed-format tests, such as NOHARM, DIMTEST, and DETECT (Svetina & Levy, 2012). Even for those procedures that appear technically feasible for mixed-format tests, their performance in this new context has not been studied as fully as would be desired in the literature, especially for large-scale operational data.

The potential problems concerning examination of the dimensionality of mixed-format tests motivated the current study. The primary goal of this study was to apply a selected set of exploratory factor analysis (EFA) methods to analyze the dimensional structures of four College Board Advanced Placement (AP) exams, all of which include an MC section and an FR section. The use of intact test forms and examinee groups could help improve knowledge of the general pattern of similarities and dissimilarities among results from different EFA methods under real settings.

Theoretical Framework

The theoretical framework is described in this section. Topics include EFA and test dimensionality, determining the number of dimensions, and exploring factor structure.

EFA and Test Dimensionality

The term EFA, in a broad sense, stands for a class of statistical procedures used to explain the observed variances and covariances, including both principal component analysis (PCA) and principal factor analysis (PFA) (Kline, 2010). However, these two procedures have different focuses and algorithms. PCA finds a group of linear combinations of observed variables, called principal components, to solve the variance maximization problem. It can be

easily done based on eigenvalue decomposition of a covariance or correlation matrix. PFA selects the smallest number of factors to adequately explain correlations among the observed variables. Factor solutions are usually obtained through maximum likelihood (ML) or least squares (LS) estimation, and orthogonal or oblique rotations are typically employed to enhance the interpretations of results. In much of the literature, a narrow definition of EFA includes PFA only. In the present chapter the term EFA is used to refer to both PCA and PFA.

From the perspective of EFA, the number of dimensions of a test is defined as the number of components or factors to retain. A test is considered to be unidimensional when only one component or factor is kept, and otherwise some degree of multidimensionality occurs. A test's internal structure corresponds to a particular factor solution produced by EFA. It should be noted that EFA does not require the use of a hypothesized dimensional structure, which seems advantageous for exploratory purposes. But if several competing hypotheses have been proposed, a confirmatory factor analysis (CFA) would be preferable to evaluate which hypothesized structure works best given the observed data, although this is beyond the scope of this chapter.

A perfectly unidimensional test, where all the items strictly measure a single dimension, is rarely seen in operational mixed-format testing. As a result, in most cases, a more important issue is whether the test is essentially unidimensional, meaning how many nontrivial dimensions are reflected by a set of items? Final decisions on test dimensionality are made based on both analytic solutions and subjective judgments.

Determining the Number of Dimensions

This study considered four types of PCA methods for determining the number of dimensions: eigenvalue rules (Hattie, 1985; Kaiser, 1960; Lord, 1980; Reckase, 1979), scree test (Cattell, 1966), minimum average partial test (MAP; Velicer, 1976), and parallel analysis (PA; Horn, 1965). Descriptions of these methods are presented here, followed by an overview of relevant research.

Eigenvalue rules. Three eigenvalue rules for PCA were used in this study as guidelines to decide whether a test is essentially unidimensional. Rule 1 keeps those principal components with eigenvalues greater than one, well known as Kaiser's eigenvalue-greater-than-one rule or K1 rule (Kaiser, 1960). Rule 2 claims a test is essentially unidimensional if at least 20% of the total variance is explained by the first principal component (Reckase, 1979). Rule 3 compares the

ratios of consecutive eigenvalues, λ_1/λ_2 and λ_2/λ_3, where λ_i indicates an eigenvalue of a correlation matrix (Hattie, 1985; Lord, 1980). If λ_1/λ_2 is three times larger than λ_2/λ_3, the test is considered to be essentially unidimensional.

Scree test. In the scree test, eigenvalues of a correlation matrix are plotted against their component numbers, and components with eigenvalues above a sharp break of the slopes of the plotted line, called an "elbow," are retained (Cattell, 1966).

MAP test. The MAP test (Velicer, 1976) is created based upon the rationale that when a sufficient number of principal components are partialled out from the original correlation matrix **R**, the resulting partial correlation matrix $\mathbf{R}^*$ should be as approximate as an identity matrix **I**, where variables no longer share any variance in common. Different MAP criteria have been considered in the original MAP test and its variants to quantify the degree of approximation between matrices $\mathbf{R}^*$ and **I**. Assuming there are *p* observed variables, the first step is to obtain the $p \times p$ partial correlation matrix $\mathbf{R}_m^* = (r_{ij,m}^*)$ for $m = 1, \ldots, (p-1)$, as

$$\mathbf{R}_m^* = \mathbf{D}^T(\mathbf{R} - \mathbf{A}_m\mathbf{A}_m^T)\mathbf{D}, \tag{1}$$

where $\mathbf{A}_m$ is the principal component loading matrix for the first m components, and **D** scales a covariance matrix into a correlation matrix. In the original MAP test, the average squared partial correlation serves as the MAP criterion, which is given by

$$\text{MAP}_m = \frac{\sum\sum {r_{ij,m}^*}^2}{p(p-1)} = \frac{\text{tr}\left(\mathbf{R}_m^{*2}\right)-p}{p(p-1)}, \tag{2}$$

where the trace, $\text{tr}(\mathbf{A})$, is the sum of diagonal elements of a squared matrix. A baseline is defined by averaging the squared correlations in $\mathbf{R} = (r_{ij})$ as

$$\text{MAP}_0 = \frac{\sum\sum r_{ij}^2}{p(p-1)} = \frac{\text{tr}(\mathbf{R}^2)-p}{p(p-1)}. \tag{3}$$

If $\text{MAP}_0 < \text{MAP}_1$, no component is retained; otherwise, the number of components is decided when the smallest MAP_m is found. Recently, Velicer, Eaton, and Fava (2000) suggested some new MAP criteria, for example, raising the power of partial correlations in Equations (2) and (3) from the second order to the fourth order, or replacing the trace of a matrix with the determinant or the largest eigenvalue.

PA procedure. PA (Horn, 1965) could be considered as a modification of K1 rule described earlier in this chapter. The K1 rule implicitly assumes that there is no sampling error, which is not true given the fact that response data are usually collected from a sample of

examinees rather than the entire examinee population (Horn, 1965). Thus, to reflect sampling error, PA begins by generating a large number of random data sets with the same number of items and examinees as the observed data. Eigenvalues of both the observed correlation and all random correlations are computed. Starting from the first position, each observed eigenvalue is compared with a certain threshold value of the distribution of random eigenvalues. Components are retained as long as the observed eigenvalue is greater than the threshold. In the original PA, random data are obtained through Monte Carlo simulation from a multivariate normal distribution, and the threshold is defined as the mean (Horn, 1965). Later, the 95th or 99th percentile threshold has been recommended by some researchers to avoid overextraction of components (Buja & Eyuboglu, 1992; Glorfeld, 1995). Regarding the generation of random data, there are also two important modifications: to replace Monte Carlo simulation with multivariate permutation from the observed data so as to skip the normality assumption (Buja & Eyuboglu, 1992), or to use regression equations to predict random eigenvalues (see Keeling, 2000 and Velicer et al., 2000 for a summary).

Overview of relevant research. Both eigenvalue rules and the scree test are easy to implement and are included in most common statistical packages such as SAS and SPSS. However, their performance with continuous variables has long been questioned. Specifically, the K1 rule has been shown to overestimate the number of dimensions in many studies (Velicer et al., 2000); Rules 2 and 3 seems vulnerable to some extreme cases (Hattie, 1985); and the process of finding an "elbow" in a scree test has been criticized because it sometimes involves too many subjective judgments (O'Connor, 2000). By contrast, MAP and PA are receiving increasing recognition that they typically produce optimal solutions, although they require relatively heavy computational efforts (O'Connor, 2000; Velicer et al., 2000). As indicated in several studies with continuous variables, MAP and PA tend to yield consistent estimation of the number of dimensions; otherwise, "the MAP tends to err (when it does err) in the direction of underextraction, whereas parallel analysis tends to err (when it does err) in the direction of overextraction" (O'Connor, 2000, p. 398).

Previous conclusions based on continuous variables are not directly applicable to the context of mixed-format tests where response data are discrete. Application of Pearson's product-moment correlations for continuous variables has been shown to distort the actual relationships among discrete variables and negatively affect EFA results, including biased

estimates of dimensions and factor loadings (Hattie, 1985; Olsson, 1979b). The polychoric correlations that describe the relationship between two continuous latent variables underlying discrete item responses have been generally considered as a substitute, but they introduce some additional problems. For example, a polychoric matrix might fail to be positive definite (PD), especially with small sample sizes. Eigenvalues of such a matrix are not always positive, so that they no longer represent variances explained by the principal components (Garrido, Abad, & Ponsoda, 2013). Several smoothing methods have been developed to transform a nonpositive definite (NPD) matrix to a PD matrix (Wothke, 1993). A practical concern might arise when smoothed polychoric correlations deviate too much from their original values.

Given limited research, the performance of MAP and PA using polychoric correlations has also not been fully understood. In a recent Monte Carlo simulation study, Garrido, Abad, and Ponsoda (2011) reported that MAP using polychoric correlations resulted in more accurate estimates of the number of dimensions than the Pearson correlations, particularly when categorical response variables had large amount of skewness. They also found a high percentage of NPD polychoric matrices as the sample size dropped down. But their findings did not support conclusions of Velicer et al. (2000) made on continuous variables that raising the partial correlation matrix to the fourth power would improve performance of the original MAP test in terms of accuracy and stability. According to some recent simulation studies on PA with polychoric correlations for discrete variables, the application of polychoric correlations also outperformed the Pearson correlations when categorical response variables showed large amount of skewness, but surprisingly PA with Pearson correlations performed at least as well as PA with polychoric correlations across multiple simulated conditions; even when PA with polychoric correlations appeared slightly superior, its effectiveness was compromised by heavy computational burden and frequent convergence issues (Cho, Li, & Bandalos, 2009; Garrido, et al., 2013; Timmerman & Lorenzo-Seva, 2011; Weng & Cheng, 2005). Due to the lack of relevant research, it is uncertain whether permutation and regression equations lead to satisfactory results for discrete variables (Garrido et al., 2013).

Exploring Factor Structure

After deciding the number of dimensions, item-level EFA on polychoric correlations could be used to explore factor structure of mixed-format tests. Specifically, Mplus (Muthén & Muthén, 1998–2012) was used in this study because it has been acknowledged as one of the most

popular programs for assessing test dimensionality and one of the most flexible packages to handle different types of data (Svetina & Levy, 2012). In Mplus, both ML and LS estimation methods are available, as well as several different types of rotations. The output file contains estimates of factor loading, R-squares, and residuals. Some model-fit statistics are also provided, including model chi-square (χ_M^2), root mean square error of approximation (RMSEA), and root mean square residual (RMSR). The first two statistics are generally available under ML and some types of LS estimation, and a smaller value indicates better fit of an EFA model. But χ_M^2 is very sensitive to large sample sizes, making itself a less preferable choice for large-scale test data (Kline, 2010). RMSR measures the overall difference between the observed correlations and correlations predicted by an EFA model, so that a smaller value represents better model fit.

Method

A detailed description of methodology is provided in this section, including data preparation, estimation and smoothing of polychoric correlations, and the use of different EFA methods to explore the number of dimensions and factor structure.

Data Preparation

Data for this study were selected from four AP exams administered in 2011, including English Language and Composition (referred to as English Language), Spanish Language and Culture (referred to as Spanish Language), Comparative Government and Politics, and Chemistry. For each exam, both the main and alternate forms were considered, which were built to the same specifications and shared a set of common MC items. The two forms were analyzed separately as cross-validation of the EFA solutions. In total, eight data sets from four exams were used in this study.

These exams represent three general subject areas, Language, Sciences, and Social Science, allowing for the investigation of whether similar dimensional structure patterns occur across different subjects. Selected data also demonstrate some unique features that potentially impact their internal structures. For example, although English Language and Spanish Language both measure language proficiency, these two language exams differ from each other in at least three aspects. First, in the samples of examinees being studied, English Language focuses on native language proficiency, whereas Spanish Language assesses mastery of a second or foreign language. Second, English Language covers only reading and writing skills, whereas Spanish Language addresses a variety of language skills including listening, reading, speaking, and

writing. The latter also contains more integrated tasks requiring examinees to combine multiple language skills, which could further complicate its dimensional structure (The College Board, 2011d). Third, only for English Language, all MC items tap reading skills, and all FR items tap writing skills. This might confound the interpretation of clusters between MC and FR items, because those clusters also reflect the difference between reading and writing skills. The use of testlets, sets of items sharing the same stimuli, has also been viewed as a potential source of multidimensionality (DeMars, 2012). For the two language exams, all the MC items are grouped into testlets; for Chemistry, only a few MC items are in testlests; for Comparative Government and Politics, all the MC items are not in testlets.

Selected exams are all mixed-format tests containing both MC and FR items. The test length varies across exams: English Language has only 55 MC items and 3 FR items, whereas Chemistry has over 80 items (75 MC items plus 6 FR items). To best fulfill the proposed test purposes, different types of FR items are also used, for instance, short answer, long answer, synthesis essay, and speaking prompts under interpersonal and presentational scenarios (The College Board, 2011a, 2011b, 2011c, 2011d). For this study, MC items were number-correct scored, and FR items were polytomously scored. Examinees who did not respond to at least 80% of the MC items were removed. A summary of test information and sample sizes for selected exams is provided in Table 1.

Operational AP exams use non-integer weights to form composite scores, keeping the MC and FR section contributions to the composite score aligned with the test specifications. However, the use of weights is not important for the EFA methods examined in this study. Thus, to simplify the computations, summed scores were used, meaning that each MC or FR point was assigned a weight of one.

It should be noted that the results and findings in this study have no direct implication for operational AP exams. First, in this study, operational data were modified in several ways for illustrative purposes. Second, as an empirical method comparison study, there is no direct way to evaluate which dimensional solution is the most accurate because the truth is never known. Further, the general accuracy of these EFA methods for mixed-format tests is not clear given very few studies in the literature. The primary focus of this study is on how different EFA methods perform and to what extent they provide consistent results, and not on the characteristics of operational AP exams.

Polychoric Correlations and Smoothing Procedure

Olsson's (1979a) "two-step" ML estimation was used to estimate polychoric correlations among all individual items. NPD matrices were smoothed by the ridge procedure, i.e., adding a small constant to diagonal elements until all the eigenvalues are positive (Wothke, 1993).

User-defined functions in R (R Core Team, 2014) were written to compute and smooth polychoric matrices for all the data sets. Specifically, the polycor package in R (Fox, 2010) was used to estimate the polychoric correlation between any pair of item score variables. Results were validated by those obtained by Mplus. The main reason to choose R was its flexibility for creating and combining user-defined functions to implement a variety of statistical procedures, which is crucial here because several EFA methods discussed in this chapter have not been included in commonly available software.

Dimensionality Assessment Using EFA

Determining the number of dimensions. Based on eigenvalues of polychoric matrices (smoothed if necessary), three eigenvalue rules (Rules 1, 2, and 3) were followed to check if essential unidimensionality held, and scree plots were created to visually reflect the relationship between any two successive eigenvalues.

For MAP, two types of correlation and two options for matrix powers were considered in this study, resulting in four combinations:

- MAP with Pearson partial correlations squared (MAP-R2),
- MAP with Pearson correlations to the fourth power (MAP-R4),
- MAP with polychoric partial correlations squared (MAP-P2), and
- MAP with polychoric partial correlations to the fourth power (MAP-P4).

For PA, only Pearson correlations were used because PA with polychoric correlations is considerably time consuming for large-scale data. Moreover, no well accepted guidelines exist in the literature regarding how to treat NPD polychoric matrices occurred in Monte Carlo simulation and permutations. PA based on regression equations was also eliminated because no equation in the literature has been proposed for the polychoric matrix. Thus, four PA procedures were included:

- PA based on Monte Carlo simulation with a mean threshold (PA-MCm),
- PA based on Monte Carlo simulation with a 95th-percentile threshold (PA-MC95),
- PA based on permutations with a mean threshold (PA-Pm), and

- PA based on permutations with a 95th-percentile threshold (PA-P95).

The number of replications in PA-MCm and PA-MC95 was 500. Comparably, the number of permutations in PA-Pm and PA-P95 was also 500, including the actual data and 499 random permutations.

Several user-defined functions in R were developed to carry out these PCA methods; R codes for MAP-R2, PA-MCm, and PA-MC95 were based on SAS macros in O'Connor (2000).

Exploring factor structure. Item-level factor analysis was conducted on polychoric matrices using Mplus. Robust weighted LS estimation was chosen for mixed-format test data. Under this estimation, values of χ^2_M, RMSEA, and RMSR were provided. Promax rotation was selected because underlying dimensions, if any, would likely be correlated to each other for a practical mixed-format test.

Model selection was made according to estimates of factor loading and factor structure, increments in R-squares, residuals, and model-fit statistics. In this study, the following guidelines were followed when choosing between models, which have also been suggested in the literature (Kline, 2010; Stone & Yeh, 2006).

First, under promax rotations, factor loadings no longer represent correlations between items and underlying dimensions, so they should be evaluated along with the corresponding structure coefficients. If both exceed 0.3, the relationship between the item and the dimension is considered to be substantial, and a dimension is considered to be nontrivial if it is substantially related to more than five items. When the dimension appears to reflect the item format effect, the five-substantial-item requirement might be loosened, because for some tests the number of FR items is less than five.

Second, R-squares of items represent the proportions of observed variances explained by the factor solution and are often expected to be large. But in EFA, all items are always assumed to load on all factors, which is likely to be inconsistent with a test's underlying structure. Consequently, the values of R-squares might be far from satisfactory under general guidelines, and instead, increments in R-squares by using (m + 1) factors than m factors might be more helpful for weighting gain-and-loss when choosing between models.

Third, residuals are expected to be small in absolute values, for instance, no greater than 0.10, and without any specific patterns. The RMSR statistic for the overall model residual is suggested to be no greater than 0.05 for an acceptable EFA model.

Fourth, compared to χ^2_M, RMSEA is more favorable and a rule of thumb is that the value of RMSEA statistic less than 0.05 indicates good-fit of the EFA model.

Last, significant tests are probably misleading given large sample sizes used in this study. Solutions with fewer dimensions and easier interpretability are preferable.

Results

Results of this study are summarized in this section. First, results from some preliminary analyses are presented. Next, numbers of dimensions decided by different PCA methods are compared. Specific dimensional solutions are then displayed separately for each test.

Results of Preliminary Analyses

Common-item effect sizes for English Language, Spanish Language, Comparative Government and Politics, and Chemistry were −0.112, 0.009, 0.178, and −0.085, respectively. These statistics indicate how groups of examinees taking the main and alternate forms differ from each other, which are calculated for alternate form minus main form common-item scores. A positive value indicates that examinees taking the alternate form are more able than those taking the main form. As Reckase (2009) emphasized, test dimensionality represents "the relationships in the data matrix that results from the interaction between a particular sample of examinees and the particular sample of items" (p. 201). For each exam, two forms were developed on the same specifications, making them parallel to a large extent. However, if two groups of examinees taking the forms were substantially different in ability, indicated by the common-item effect size, a discrepancy in the dimensional structure between two forms might be present. For Spanish Language, the difference between two groups taking the main and alternate forms was almost unnoticeable. For English Language and Chemistry, the differences were around 0.1 in magnitude and might be considered to be small. For Comparative Government and Politics, the difference was nearly 0.2, which was relatively large compared to the other three exams.

Reliability coefficients using Cronbach's alpha (Cronbach, 1951) were estimated separately for MC and FR sections, as presented in Table 2. The FR section typically had lower reliability than the MC section. Disattenuated MC and FR correlations were calculated to roughly evaluate whether MC and FR items measure essentially the same dimensions, which are also provided in Table 2. A value close to one indicates that combining MC and FR items is unlikely to introduce extra dimensions. The disattenuated correlations for English Language

were the lowest, approximate 0.80 for both the main and alternate forms, suggesting that MC and FR items might measure somewhat different dimensions. By contrast, the disattenuated correlations for Chemistry exceeded 0.97 for both forms; separate dimensions associated with item formats might not be necessary. In general, the disattenuated correlations for the two language exams were lower than those for Comparative Government and Politics and Chemistry.

The polychoric matrices were PD for all the data sets except for the Comparative Government and Politics alternate form. The sample size for this form was only 618, whereas the other data sets involved more than 3,000 examinees. Eigenvalues of both the original and smoothed polychoric matrices for this form are displayed in Figure 1. It seemed that smoothing by adding a ridge did not change the general pattern of eigenvalues.

Estimated Number of Dimensions

Results of eigenvalue rules. Results of three eigenvalue rules based on PCA with polychoric correlations presented a mixed picture. In Table 3, all the data sets contained more than one dimension according to Rule 1 (K1 rule). Results of Rule 2 suggested exactly the opposite: all the data sets were considered to be essentially unidimensional, as the proportions of total variance were all greater than 20%. Under Rule 3, Spanish Language was not essentially unidimensional because the ratio of λ_1/λ_2 to λ_2/λ_3 did not reach three, but the other three exams were considered to be essentially unidimensional.

Scree plots. Scree plots in Figures 2 through 5 were examined for the occurrence of multidimensionality in the data. For each exam, the plots for the main and alternate forms looked almost identical except for English Language. Even for English Language, the overall pattern of the difference between every two consecutive eigenvalues was still very similar across forms. The two language exams, especially Spanish Language, showed a higher degree of multidimensionality compared to Comparative Government and Politics and Chemistry.

Results of MAP and PA. Results of four MAP tests (MAP-R2, MAP-R4, MAP-P2, and MAP-P4) and four PA procedures (PA-MCm, PA-MC95, PA-Pm, and PA-P95) are summarized in Table 4. Among four MAP tests, MAP-R2 always was associated with the smallest number of dimensions, whereas the other three MAP tests tended to give similar results. The four different PA procedures, however, generally suggested the same number of dimensions reflected by each data set.

Two major findings were made about the comparison of MAP and PA methods. On the one hand, both MAP and PA methods indicated that every data set exhibited some multidimensionality. On the other hand, the numbers of dimensions estimated by MAP and PA methods were inconsistent for most of the data sets except Comparative Government and Politics. The results support some previous findings that MAP tests tend to suggest fewer dimensions than PA, but sometimes the difference in the number of dimensions estimated by two types of methods was too large to make any definite conclusions.

Figures 6 through 9 contain all the PA plots allowing some closer inspections of PA results. It shows that the first two or three eigenvalues seemed significantly above the threshold value, but the next several eigenvalues were very close to the corresponding threshold values, and then at a point, observed eigenvalues began to fall below the threshold values. The PA procedure retained those components with eigenvalues greater than a certain threshold regardless of distance. Components with eigenvalues slightly above the threshold values might be trivial.

Estimated Factor Structure

When using Mplus, the range of number of factors to be considered needed to be imputed before conducting EFA. In this study, the minimum number of factors was always set to one, implying that the unidimensionality held. The maximum number of factors was decided based on results of PCA methods discussed in the previous section. For PA, the PA plots in Figures 6 through 9 were also used to delete seemingly trivial dimensions. Specifically, the maximum numbers of factors for the four exams were assigned in a conservative manner: three for Comparative Government and Politics, four for English Language and Chemistry, and five for Spanish Language.

Additionally, polychoric correlations estimated by Mplus were slightly different from those estimated by R, probably because (a) Mplus treated a discrete variable with more than ten response categories as a continuous variable, and (b) Mplus did not smooth any NPD polychoric matrices. Nevertheless, eigenvalues and EFA solutions did not seem to be greatly affected.

For English Language, Spanish Language, and Chemistry, only the EFA solutions for the main form are presented here, because similar patterns of factor structure held for the alternate form. For Comparative Government and Politics, the EFA solution for the main form is presented followed by a brief description of the alternate form result where some differences were found.

English Language. Four EFA models (1-factor, 2-factor, 3-factor, and 4-factor models) were examined for English Language main form data. In Table 5, the values of RMSEA and RMSR statistics for the four models fell below 0.05, indicating that the four solutions all fit the data reasonably well and the overall residuals were acceptable. Among them, the most parsimonious 1-factor model might be preferred. A quick inspection of factor loadings and factor structures of each solution revealed that 3-factor and 4-factor models contained some dimensions that were associated with only three items with substantial pattern coefficients, but whether they were trivial dimensions should be decided combining other results.

A residual analysis partly confirmed the previous findings but provided further support for examining a 3-factor solution. For a 1-factor model, the majority of residuals were between −0.10 and 0.10, suggesting that the correlations estimated by a 1-factor model were close to those observed in the data. However, some residuals were still around or greater than 0.15 in absolute value, and two special patterns were found. First, the correlations among MC items from a testlet (MC42–MC54) tended to be underestimated, whereas the correlations between these items and the rest of items tended to be overestimated. Second, the correlations among three FR items were underestimated. Only after two additional factors were added, all the residuals were considerably small and such special patterns became largely weakened. Table 6 presents a simplified 3-factor solution keeping the substantial pattern coefficients only.

Results of R-squares using different factor models are plotted for individual items in Figure 10. As seen in the figure, the observed variances for the testlest of MC42 to MC54 and for three FR items were better explained as the second and third factors adding to a 1-factor model, respectively, which also provide some evidence that a 3-factor model might fit the data better.

In sum, for English Language, essential unidimensionality might be assumed according to the model-fit statistics. However, a 3-factor model might be more useful for understanding the test's internal structure. A plausible interpretation of this 3-factor model is provided in the Discussion section.

Spanish Language. Five EFA models (1-factor, 2-factor, 3-factor, 4-factor, and 5-factor models) were examined for the Spanish Language main form data. In Table 7, the value of RMSR for a 1-factor model exceeded 0.05, suggesting that the use of a single factor might be insufficient for depicting correlations between observed item responses. When a 2-factor model was used, the value of RMSR reduced by half, from 0.061 to 0.032, and RMSEA continued to

fall from 0.039 to 0.023, both implying that a multidimensional model might fit the observed data better than a unidimensional model. An examination of the pattern in coefficients and residuals further showed that models having four or five factors were not interpretable. For example, negative factor correlations occurred in a 5-factor solution, which is a typical sign of overfitting. Additionally, after three factors were included in the model, most of the residuals were between −0.10 and 0.10 and no special pattern was found; introducing any extra factors into the model did not significantly improve the model-fit. As a result, either a 2-factor or 3-factor model might be considered for Spanish Language main form data.

A 3-factor model might be more favorable than a 2-factor model because some residuals of a 2-factor model were still relatively large in absolute value. Specifically, the intercorrelations within three testlests, MC1–MC4, MC53–MC61, and MC62–MC70, respectively, appeared to be underestimated. The use of three factors reduced the residual magnitudes and improved the overall model-fit. Compared to a 2-factor solution, a 3-factor solution was also less confusing, and slightly fewer items did not load on any factor under this solution (14/74 = 19%). Table 8 presents a simplified 3-factor solution identifying only the substantial pattern coefficients for each factor.

Results of R-squares of a 1-factor model and R-square increments of models having two to five factors are shown in Figure 11. It can be seen that compared to a 1-factor model, using a 2-factor model better predicted the observed variances, and adding one additional factor (3-factor model) further improved the prediction of variances for several items (e.g., MC13, MC44, MC69, and FR3). However, relations between the dimensions and the language skills measured (i.e., reading, listening, writing, and speaking skills) were still unclear.

In sum, for Spanish Language, some degree of multidimensionality was detected. EFA results suggested a 2-factor or 3-factor model, but further information would still needed to make the final decision, for instance, the purpose of the dimensionality assessment.

Comparative Government and Politics. Three EFA models (1-factor, 2-factor, and 3-factor models) were investigated for Comparative Government and Politics main form data. In Table 9, when a 1-factor model was used, the value of RMSEA was smaller than 0.020, and RMSR was approximately 0.030. Although the use of two or three factors resulted in much smaller values of RMSEA and RMSE, the most parsimonious 1-factor model seemed satisfactory in describing the internal structure of this set of data, and the essential

unidimensionality could be assumed. An examination of residuals also provided strong evidence for a 1-factor model. No special pattern was found, and the average residual magnitude was approximately 0.023. Correlations predicted by a 1-factor model were similar to those between observed variables. Results of R-squares in Figure 12 again showed a preference for a 1-factor model: using a multidimensional model did not seem to explain considerably more variance in most of the observed variables than a 1-factor model.

The same ranges of EFA models, from 1-factor to 3-factor models, were investigated for Comparative Government and Politics alternate form data. Similarly, the values of RMSEA for all three models were smaller than 0.020, so that a 1-factor model might be preferred. However, the value of RMSR for a 1-factor model was nearly 0.070, and after two more dimensions were included (3-factor model) RMSR was still as high as 0.057, indicating that the overall discrepancy between observed and estimated correlations was more than acceptable when the EFA model had three or less factors. Specific detection of residuals partly confirmed the result of RMSR: there were many residuals greater than 0.10 in absolute values for 1-factor model, although it was not obvious if any special pattern existed.

In short, given factor analysis results alone, it was unclear whether essential unidimensionality could be assumed for Comparative Government and Politics alternate form data. Further, such difference between the main and alternate form data were somewhat unexpected, because PCA methods for determining the number of dimensions produced relatively consistent results across different forms for this exam, as shown in Tables 3 and 4 as well as Figures 4 and 8. Potential explanations included the small sample size and NPD polychoric matrix used in factor analysis for the alternate form. This might result in large sampling error and bias that negatively affect the EFA solutions. The difference in ability between groups taking the main and alternate forms, indicated by the common-item effect size, might also be considered due to the sample specific nature of test dimensionality.

Chemistry. Four EFA models (1-factor, 2-factor, 3-factor, and 4-factor models) were explored for Chemistry main form data. As shown in Table 10, the values of RMSEA and RMSR for all four models were acceptable under the 0.05 criteria. Specifically, RMSEA and RMSR were as small as 0.022 and 0.031 respectively for a 1-factor model. Their values kept falling when more dimensions were added, but marked reductions only occurred when a 2-factor model was used instead of a 1-factor model (approximately 18% and 29% reduction in RMSEA and

RMSR, respectively). Patterns of the coefficients and residuals also implied that models with one or two factors might be desired, although there was no clear preference for either one of them.

Results of R-squares and R-square increments are displayed in Figure 13. As seen in the figure, improvements in the estimation of the observed variance made by using a multidimensional model did not seem to be salient, except for only a few items such as MC37 and MC59.

Combining the EFA results together for Chemistry main form data, essential unidimensionality might be assumed. A 2-factor model appeared to be slightly better in estimating the underlying correlations between variables, but model complexity and further issues related to the use of a multidimensional model should also be taken into consideration.

Discussion

The purpose of this study was to empirically investigate how different EFA methods performed for examining the dimensional structure of mixed-format tests. Data from four AP exams, English Language, Spanish Language, Comparative Government and Politics, and Chemistry, were used as illustrative examples. Selection of data was intended to cover a broad range of subject areas and represent various test and sample features. Specific methods included three eigenvalue rules, scree test, MAP and PA procedures, and model selection based on several EFA criteria.

By using real mixed-format test data, this study allowed some practical challenges in examining the test dimensionality to emerge. For example, theoretically it seems more appropriate to conduct EFA on polychoric correlations than on Pearson correlations, but in practice the estimation of polychoric correlations might be problematic. With large sample sizes and test lengths, the estimation of polychoric correlations usually requires a considerable amount of time and computing resources. This might make some simulation-based methods less feasible, such as several PA procedures. However, when the sample size is small, the estimated polychoric correlations contain substantial large sampling error and the polychoric matrix might be NPD. This also hinders subsequent dimensionality analyses. This chapter used a simple procedure to smooth NPD matrices, and results showed that the smoothed PD matrix was similar to the original one. Further studies are needed to evaluate the benefits of using polychoric correlations under different conditions for mixed-format tests.

In addition, there is one technical limitation regarding Mplus: for MC items, guessing behavior was not modeled when Mplus was used to explore a test's dimensional structure. However, Mplus might be the only well-established software program that is currently available for conducting EFA with polychoric correlations for mixed-format test data.

Despite the lack of knowledge of the true dimensional structure for each test, some general patterns of similarities and dissimilarities among results from different EFA methods were observed in this study. Specifically, Rule 1 (K1 rule) tended to identify more dimensions compared to other methods, which supports conclusions of previous studies (e.g., Velicer et al., 2000). When the purpose of the dimensionality assessment was to check whether essential unidimensionality held, Rules 2 and 3 provided almost identical results to those of a more complicated factor analysis. Recall that according to Rules 2 and 3, English Language, Comparative Government and Politics, and Chemistry were all considered to be essentially unidimensional. Such results were later confirmed by EFA model selection. For Spanish Language, Rule 3 and EFA model selection reached agreement in the presence of some degree of multidimensionality. Rule 2 still considered this exam to be unidimensional, but the percentages of variance explained by the first principal component were close to the 20% criterion (22% and 24% for the main and alternate forms, respectively), which might also indicate some tendency towards multidimensionality in these data sets. Results of MAP and PA were consistent with some previous studies that dimensions decided by MAP are often smaller than those by PA (O'Connor, 2000). However, both procedures appeared to be "sensitive" to the occurrence of trivial dimensions. Especially for PA, it would be better to interpret analytic results along with the PA plots in Figures 6 through 9. Similar to PA plots, scree plots were also found to be useful in deciding the number of dimensions when combined with results of other methods. There is still a need for method comparison studies based on operational mixed-format test data as well as simulations where certain test and sample characteristics could be manipulated.

When a factor solution is found to be preferable according to a series of statistical criteria, further information is always needed in order to make meaningful interpretations. Specific interpretations of the factor solutions are not included in this chapter, but a possible explanation of the 3-factor model for English Language is discussed here as an example. Recall that for English Language, all the MC items are grouped into testlets. Specifically, the main form is comprised of four reading testlets: Testlet 1 (MC1–MC11), Testlet 2 (MC12–MC25),

Testlet 3 (MC26–MC41), and Testlet 4 (MC42–MC54). The reading passages for the first three testlests are selected from works of 20th century literature, whereas the passage for the fourth testlet was written in the 19th century (pre-20th century literature). Such distribution of the content domains might result in a factor structure in Table 6: Factor 1 was mainly measured by MC items from Testlets 1, 2, and 3, and Factor 2 was measured by MC items from Testlet 4. More interestingly, Factor 3 was only substantially associated with three FR items. This might inform that the item format effect occurred. But given the particular test structure of English Language, Factor 3 might also reflect the difference between reading and writing skills. Another possible explanation is speededness, as these FR items are placed near the end of the test. Future studies could focus on the interpretations and explanations of certain factor solutions for AP exams or other operational mixed-format tests.

References

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*, 509–540.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets*. Unpublished doctoral dissertation, University of Maryland.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Cho, S-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*, 748–759.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*, 104–121.

Fox, J. (2010). *polycor: Polychoric and Polyserial Correlations*. R package version 0.7-8. http://CRAN.R-project.org/package=polycor

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, *71*, 551–570.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454–474.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377–393.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Keeling, K. B. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research*, *35*, 457–468.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

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

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*, 396–402.

Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, 485–500.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Stone, C. A., & Yeh, C-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement*, *66*, 193–214.

Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, *36*, 659–669.

The College Board. (2011a). *Student performance Q&A: 2011 AP® Chemistry free-response questions*. Retrieved from http://apcentral.collegeboard.com/apc/public/repository/ap11_chemistry_qa.pdf

The College Board. (2011b). *Student performance Q&A: 2011 AP® Comparative Government and Politics free-response questions*. Retrieved from http://apcentral.collegeboard.com/apc/public/repository/ap11_comp_go_po_qa.pdf

The College Board. (2011c). *Student performance Q&A: 2011 AP® English Language and Composition free-response questions*. Retrieved from http://apcentral.collegeboard.com/apc/public/repository/ap11_english_language_qa.pdf

The College Board. (2011d). *Student performance Q&A: 2011 AP® Spanish Language free-response questions*. Retrieved from http://apcentral.collegeboard.com/apc/public/repository/ap11_spanish_language_qa.pdf

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*, 209–220.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Boston, MA: Kluwer Academic Publishers.

Weng, L-J., & Cheng, C-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*, 697–716.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.

Table 1

Test Information and Sample Sizes

Exam	Form	MC Items	FR Items	FR Points	*N*
English Language	Main	54	3	9 each	401,763
	Alt	54	3	9 each	5,110
Spanish Language	Main	70	4	5 each	117,242
	Alt	70	4	5 each	3,214
Comp Government & Politics	Main	55	4	15, 5, 7, 8	16,486
	Alt	54	4	13, 4, 5, 7	618
Chemistry	Main	75	6	10, 9, 10, 15, 8, 8	114,803
	Alt	75	6	10, 10, 9, 15, 8, 9	3,540

Note. For the Comparative Government & Politics Exam (the main and alternate forms), the first FR item in this table is the sum score of the first five FR items in operational data.

Table 2

Section Reliability Coefficients, and Observed and Disattenuated MC and FR Correlations

Exam	Form	Reliability		Correlation	
		MC	FR	Obs.	Dis.
English Language	Main	0.896	0.719	0.646	0.805
	Alt	0.931	0.780	0.698	0.818
Spanish Language	Main	0.898	0.678	0.654	0.839
	Alt	0.906	0.743	0.707	0.862
Comp Government & Politics	Main	0.882	0.821	0.798	0.937
	Alt	0.899	0.709	0.716	0.898
Chemistry	Main	0.930	0.894	0.888	0.973
	Alt	0.934	0.919	0.903	0.975

Table 3

Results of Three Eigenvalue Rules

Exam	Form	# of Eigenvalues Greater Than 1	% Variance Explained by the 1st Eigenvalue	Ratio of Eigenvalues		$\frac{\lambda_1/\lambda_2}{\lambda_2/\lambda_3}$
				λ_1/λ_2	λ_2/λ_3	
English Language	Main	7	27.383	6.001	1.889	3.177
	Alt	6	36.247	9.069	1.479	6.130
Spanish Language	Main	9	22.462	3.716	2.045	1.817
	Alt	15	24.279	3.769	2.247	1.678
Comp Government & Politics	Main	7	27.499	8.460	1.323	6.394
	Alt	16	27.293	6.512	1.320	4.932
Chemistry	Main	8	30.240	9.924	1.562	6.355
	Alt	13	30.811	11.468	1.237	9.273

Note. For the Comparative Government and Politics Exam 2011 alternate form, results are derived from the smoothed polychoric matrix.

Table 4

Results of Four MAP Tests and Four PA Procedures

Exam	Form	MAP				PA			
		R2	R4	P2	P4	MCm	MC95	Pm	P95
English Language	Main	2	2	2	3	7	7	7	7
	Alt	2	3	3	6	4	3	4	3
Spanish Language	Main	3	5	5	5	9	9	9	9
	Alt	3	5	5	5	5	5	5	5
Comp Government & Politics	Main	2	2	2	3	3	3	3	3
	Alt	2	3	2	3	3	3	3	3
Chemistry	Main	2	4	4	4	7	7	7	7
	Alt	2	3	3	4	4	4	4	4

Note. For the Comparative Government and Politics Exam 2011 alternate form, results of MAP are derived from the smoothed polychoric matrix.

Table 5

Values of Model-Fit Statistics, Number of Substantial Pattern Coefficients, and Factor Correlations for English Language Exam Main Form Data

Model	χ^2_M	df_M	P-Value	RMSEA	RMSR
1-factor	742576.736	1539	0.000	0.035	0.046
2-factor	178229.639	1483	0.000	0.017	0.020
3-factor	113019.212	1428	0.000	0.014	0.019
4-factor	79343.007	1374	0.000	0.012	0.015

# of Substantial Pattern Coefficients

1-factor	2-factor		3-factor			4-factor			
1	1	2	1	2	3	1	2	3	4
51	43	12	40	12	3	16	17	3	11

Factor Correlations

1-factor	2-factor		3-factor			4-factor			
1	1	2	1	2	3	1	2	3	4
1									
2	0.590		0.582			0.731			
3			0.536	0.366		0.526	0.517		
4						0.497	0.590	0.427	

Table 6

A Correlated 3-Factor Solution for English Language Exam Main Form Data

Item	3-Factor Solution			Item	3-Factor Solution		
	Factor 1	Factor 2	Factor 3		Factor 1	Factor 2	Factor 3
MC5	0.673			MC10	0.416		
MC34	0.654			MC25	0.411		
MC36	0.636			MC38	0.408		
MC35	0.620			MC19	0.402		
MC6	0.616			MC11	0.400		
MC32	0.603			MC23	0.396		
MC4	0.572			MC20	0.373		
MC31	0.554			MC41	0.356		
MC29	0.553			MC12	0.319		
MC33	0.551			MC14	0.317		
MC7	0.545			MC15	—		
MC2	0.537			MC16	—		
MC40	0.536			MC18	—		
MC8	0.524			MC27	—		
MC1	0.517			MC24	—		
MC9	0.516			MC30	—		
MC39	0.512			MC53		1.004	
MC26	0.511			MC52		0.912	
MC45	0.504			MC54		0.782	
MC3	0.487			MC50		0.664	
MC28	0.468			MC51		0.531	
MC13	0.466			MC48		0.520	
MC17	0.464			MC49		0.465	
MC21	0.455			MC47	0.359	0.408	
MC46	0.453	0.354		MC44		0.373	
MC42	0.449	0.344		FR2			0.509
MC37	0.448			FR3			0.478
MC43	0.445	0.317		FR1			0.447
MC22	0.421						

Note. Only substantial pattern coefficients are displayed. If an item is not substantially associated with any factor, the highest pattern coefficient is kept and denoted by a dash.

Table 7

Values of Model-Fit Statistics, Number of Substantial Pattern Coefficients, and Factor Correlations for Spanish Language Exam Main Form Data

Model	χ^2_M	df_M	P-Value	RMSEA	RMSR
1-factor	479035.491	2627	0.000	0.039	0.061
2-factor	155956.036	2554	0.000	0.023	0.032
3-factor	92406.236	2482	0.000	0.018	0.024
4-factor	58203.504	2411	0.000	0.014	0.019
5-factor	39278.521	2341	0.000	0.012	0.016

# of Substantial Pattern Coefficients

1-factor	2-factor		3-factor			4-factor				5-factor				
1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
67	32	26	23	15	26	23	24	8	9	23	24	0	8	9

Factor Correlations

1-factor	2-factor		3-factor			4-factor				5-factor				
1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
1														
2	0.591		0.597			0.558				0.546				
3			0.573	0.470		0.539	0.487			-0.034	-0.234			
4						0.513	0.336	0.451		0.547	0.479	0.064		
5										0.513	0.335	-0.055	0.459	

Table 8

A Correlated 3-Factor Solution for Spanish Language Exam Main Form Data

Item	3-Factor Solution			Item	3-Factor Solution		
	Factor 1	Factor 2	Factor 3		Factor 1	Factor 2	Factor 3
MC49	0.698			MC62	0.305	0.427	
MC44	0.653			MC65		0.408	
MC14	0.619			MC59		0.386	
MC48	0.582			MC55		0.378	
MC51	0.567			MC63		0.374	
MC33	0.506			MC53		0.370	
MC46	0.478			MC57		—	
MC26	0.472			MC60		—	
MC42	0.469			MC13			0.864
MC15	0.465			FR3			0.864
MC50	0.462			MC12			0.774
MC1	0.426			MC20			0.714
MC34	0.415			MC23			0.690
MC2	0.390			MC38			0.690
MC29	0.382			MC10			0.675
MC30	0.362			MC8			0.652
MC6	0.357			MC24			0.644
MC45	0.354			MC11			0.643
FR2	0.330			MC40			0.607
MC7	0.327			MC37			0.598
MC52	0.321			MC36			0.555
MC27	0.306			MC39			0.515
MC47	—			MC43			0.466
MC5	—			FR4			0.436
MC18	—		—	MC22			0.435

MC9	—			MC54	0.311	0.431
MC67	—			MC41		0.415
MC3	—			FR1		0.406
MC32	—		—	MC17		0.381
MC16	—			MC21		0.356
MC69		0.765		MC56	0.341	0.342
MC64		0.762		MC25		0.329
MC68		0.707		MC4		0.318
MC66		0.590		MC19		—
MC61		0.539		MC31		—
MC70		0.447		MC28		—
MC58		0.439	0.398	MC35		—

Note. Only substantial pattern coefficients are displayed. If an item is not substantially associated with any factor, the highest pattern coefficient is kept and denoted by a dash.

Table 9

Values of Model-Fit Statistics, Number of Substantial Pattern Coefficients, and Factor Correlations for Comparative Government and Politics Exam Main Form Data

Model	χ^2_M	df_M	P-Value	RMSEA	RMSR
1-factor	10453.034	1652	0.000	0.018	0.030
2-factor	5606.271	1594	0.000	0.012	0.022
3-factor	4087.801	1537	0.000	0.010	0.018

# of Substantial Pattern Coefficients

1-factor	2-factor		3-factor		
1	1	2	1	2	3
50	30	19	28	17	6

Factor Correlations

1-factor	2-factor		3-factor		
1	1	2	1	2	3
1					
2	0.667		0.571		
3			0.539	0.542	

Table 10

Values of Model-Fit Statistics, Number of Substantial Pattern Coefficients, and Factor Correlations for Chemistry Main Form Data

Model	χ^2_M	df_M	P-Value	RMSEA	RMSR
1-factor	184263.276	3159	0.000	0.022	0.031
2-factor	116248.055	3079	0.000	0.018	0.022
3-factor	90419.261	3000	0.000	0.016	0.019
4-factor	68436.565	2922	0.000	0.014	0.016

# of Substantial Pattern Coefficients

1-factor	2-factor		3-factor			4-factor			
1	1	2	1	2	3	1	2	3	4
78	38	40	26	29	17	28	25	4	9

Factor Correlations

1-factor	2-factor		3-factor			4-factor			
1	1	2	1	2	3	1	2	3	4
1									
2	0.679		0.595			0.565			
3			0.559	0.603		0.556	0.566		
4						0.551	0.601	0.526	

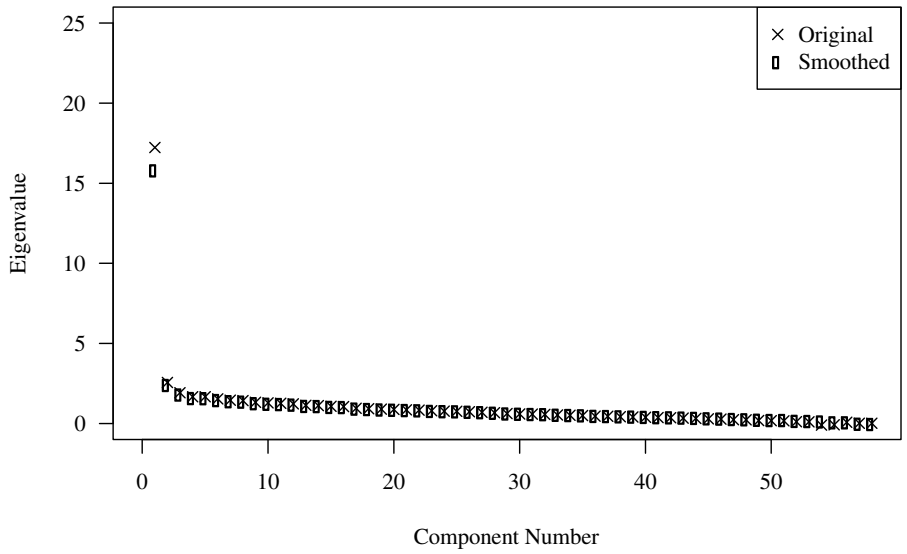


Figure 1. Eigenvalues of the original and smoothed polychoric matrices for Comparative Government and Politics alternate form.

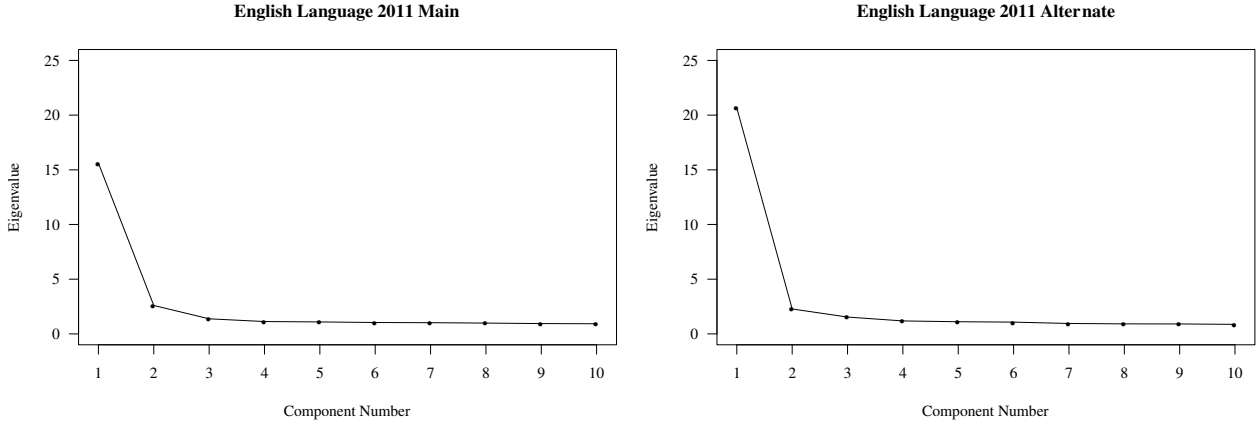


Figure 2. Scree plots of the first ten eigenvalues for English Language.

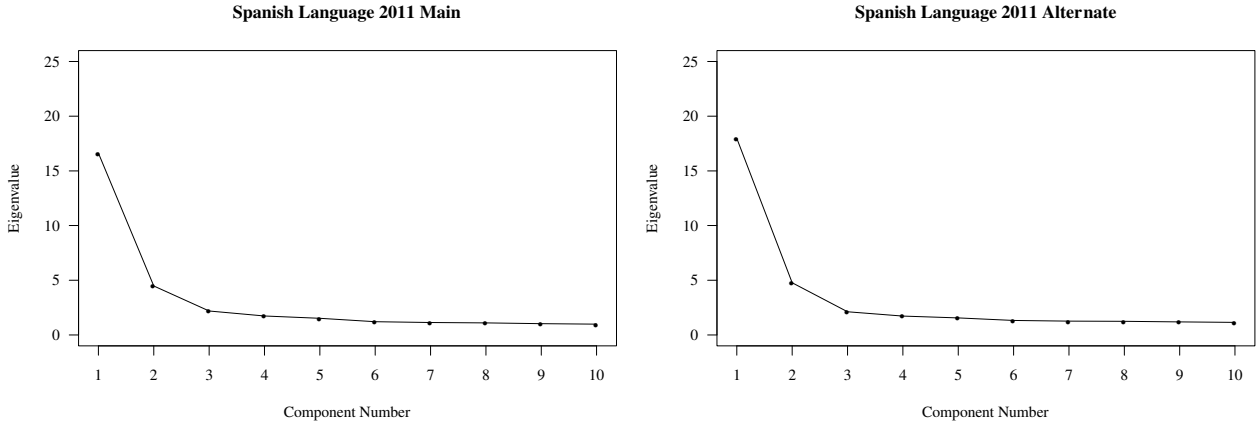


Figure 3. Scree plots of the first ten eigenvalues for Spanish Language.

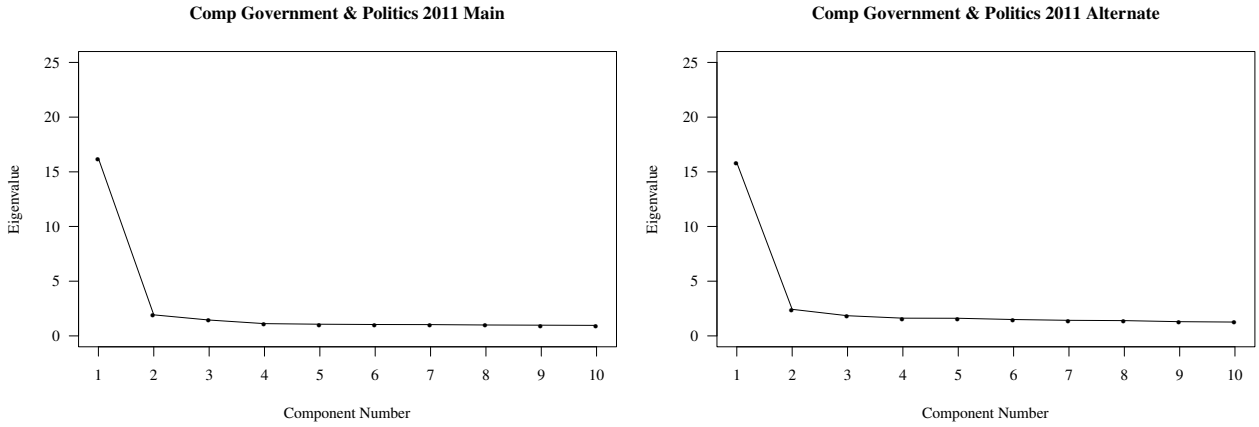


Figure 4. Scree plots of the first ten eigenvalues for Comparative Government and Politics.

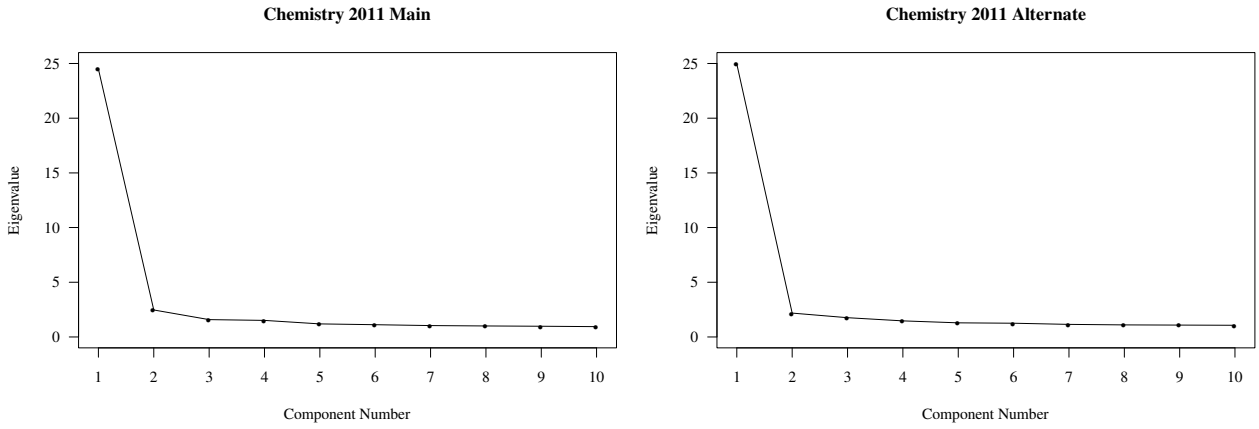


Figure 5. Scree plots of the first ten eigenvalues for Chemistry.

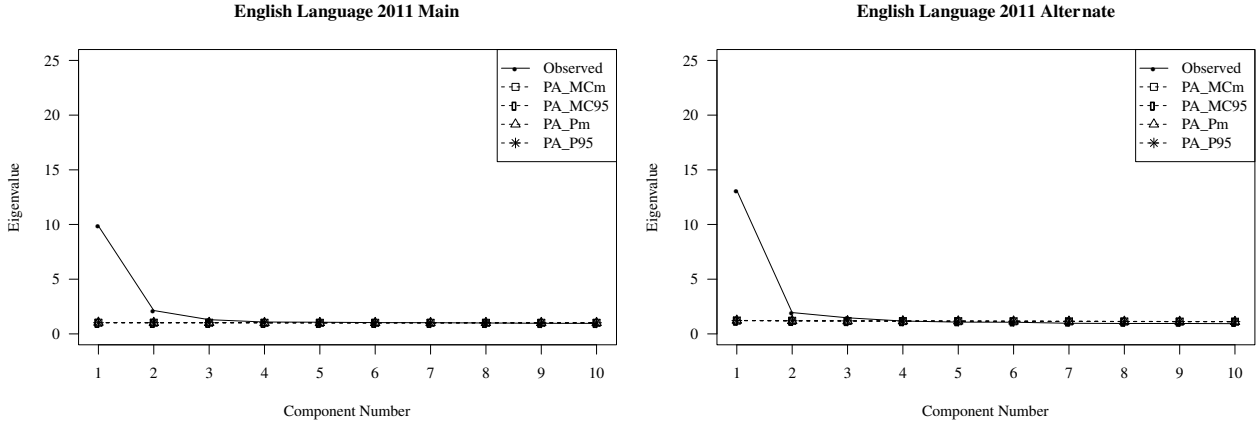


Figure 6. PA plots of the first ten eigenvalues for English Language.

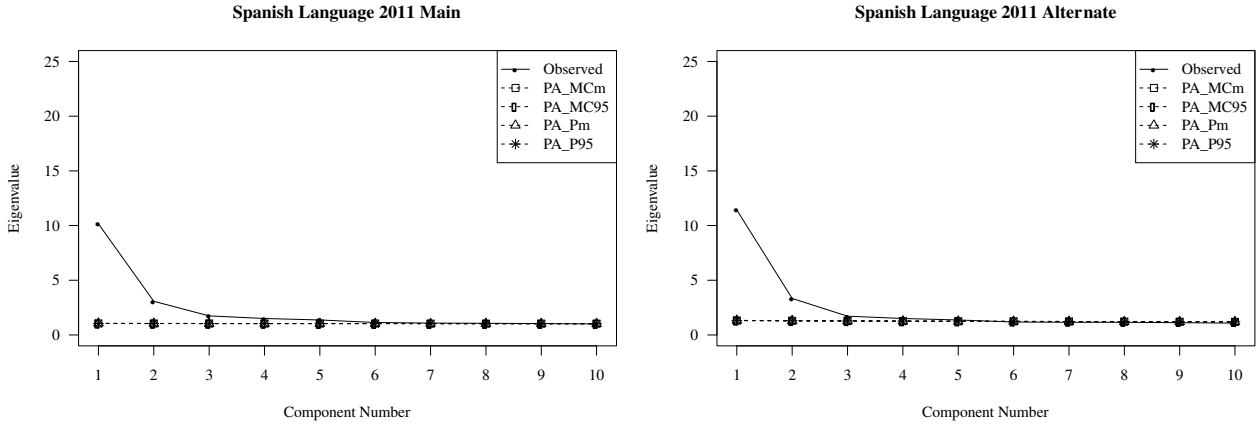


Figure 7. PA plots of the first ten eigenvalues for Spanish Language.

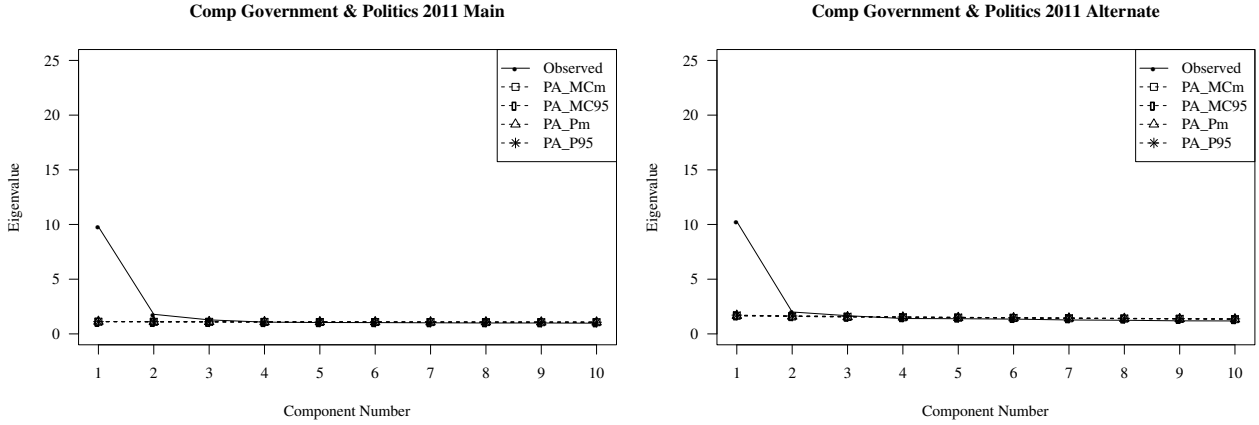


Figure 8. PA plots of the first ten eigenvalues for Comparative Government and Politics.

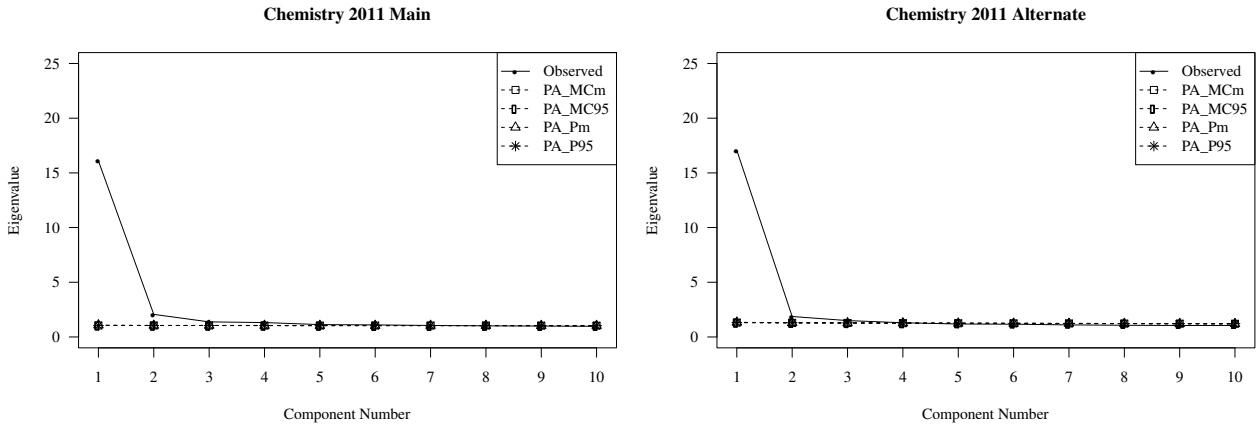


Figure 9. PA plots of the first ten eigenvalues for Chemistry.

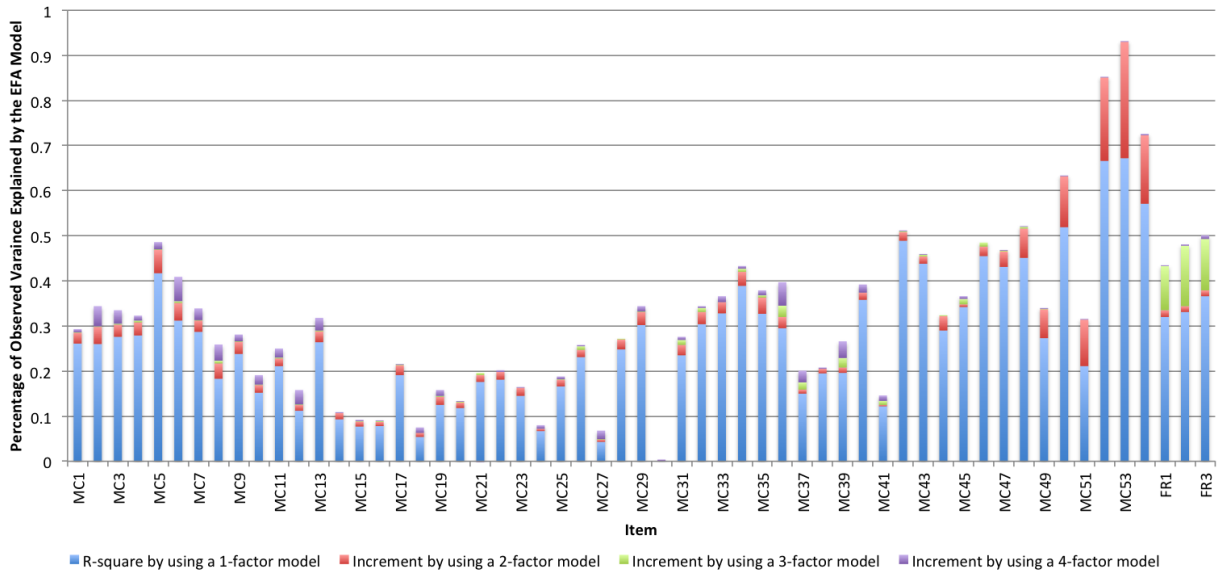


Figure 10. R-squares when using the unidimensional model and increments in R-squares when using multidimensional models for English Language Exam main form data.

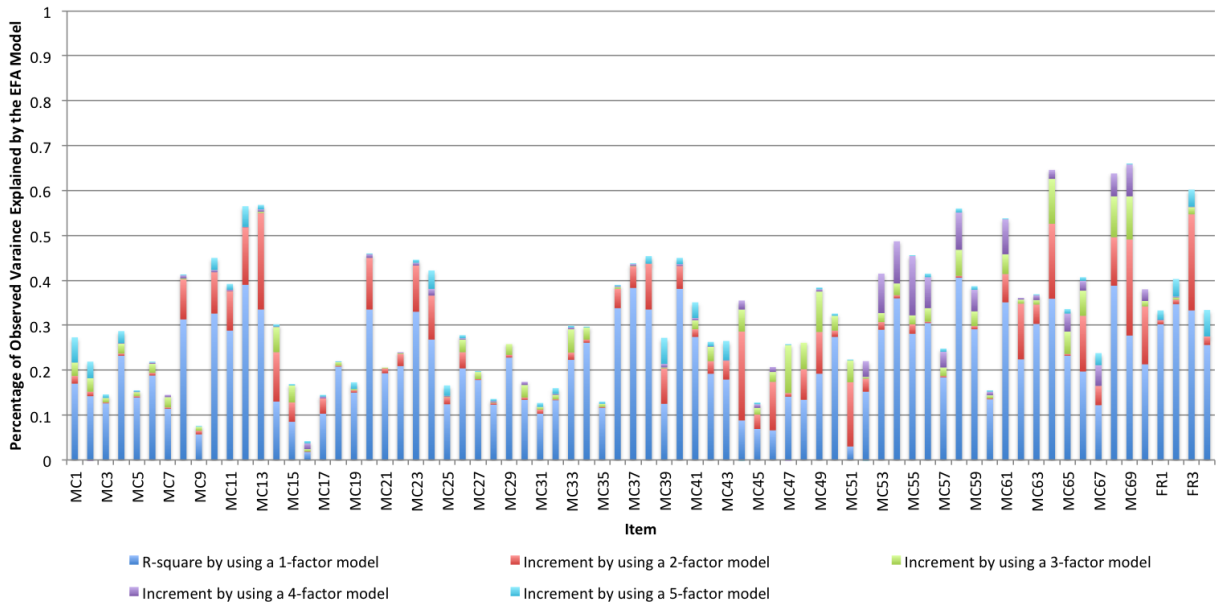


Figure 11. R-squares when using the unidimensional model and increments in R-squares when using multidimensional models for Spanish Language Exam main form data.

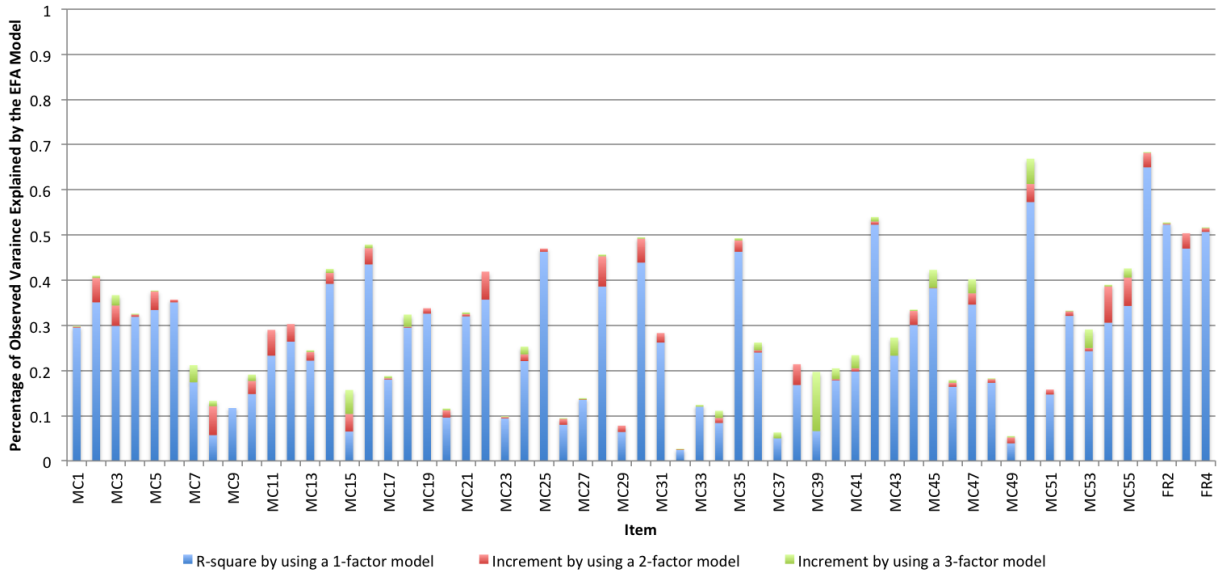


Figure 12. R-squares when using the unidimensional model and increments in R-squares when using multidimensional models for Comparative Government and Politics Exam main form data.

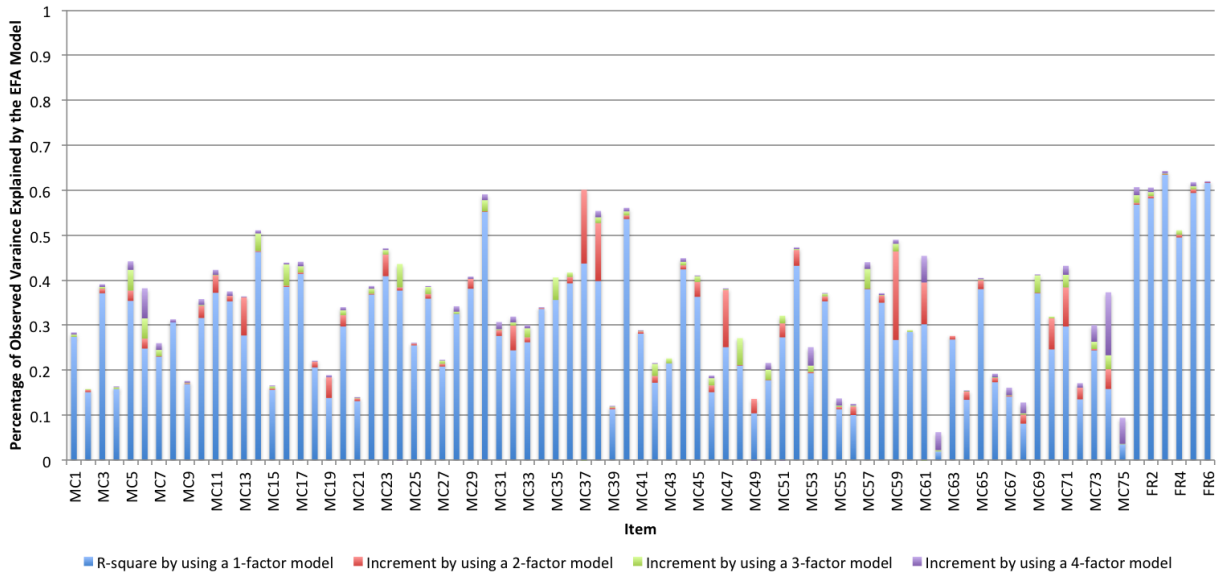


Figure 13. R-squares when using the unidimensional model and increments in R-squares when using multidimensional models for Chemistry Exam main form data.

Chapter 7: A Comparison of Unidimensional IRT and Bi-factor Multidimensional IRT Equating for Mixed-Format Tests

Guemin Lee
Yonsei University, Seoul, Korea
Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

The main purposes of this study were to develop bi-factor multidimensional item response theory (BF-MIRT) observed-score equating procedures for mixed-format tests and to investigate the relative appropriateness of the proposed procedures. Data for this study were from the College Board Advanced Placement (AP) examinations and three data sets were developed: matched samples, pseudo forms, and simulated data sets. Very minor within-format residual dependence in mixed-format tests was found after controlling for the influence of the primary general factor. Either the unidimensional IRT (UIRT) or BF-MIRT observed-score equating method can be considered for practical equating. The BF-MIRT equating method was found to produce more accurate equating results for mixed-format tests compared to the UIRT method with non-negligible multidimensionality. In this case, the BF-MIRT model fits the data better than the UIRT model. When a BF-MIRT model is implemented, we recommend the use of the observed-score equating instead of true-score equating because the latter requires an arbitrary unidimensional approximation or reduction process.

A Comparison of Unidimensional IRT and Bi-factor Multidimensional IRT Equating for Mixed-Format Tests

Over the past two decades, the use of free-response (FR) items as alternatives or supplements to traditional multiple-choice (MC) items has been one of the central features of current testing practices (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Baxter, & Gao, 1993). The FR items have been broadly used in large scale testing programs with or without MC items. Tests containing both MC and FR items are often called mixed-format tests in the literature.

One of the important psychometric practices related to the use of tests is "equating" to achieve score comparability over multiple forms (Kolen & Brennan, 2004). In the context of applying item response theory (IRT) to achieve score comparability, it would be helpful in some cases to differentiate "linking" and "equating" even though some IRT experts tend to use the two terms interchangeably. Linking (sometimes called "scale linking," which hereafter is used in this paper) refers to placing item and ability parameter estimates on the same ability scale. If item parameter estimates from different test forms are on the same scale, ability scores derived from those parameter estimates over different test forms are comparable and can be used interchangeably without any further transformations. However, when number-correct scoring is used rather than the IRT-ability scoring and a pre-developed number-correct-to-scale score conversion table is used, additional "equating" procedures, in addition to "scale linking," should be conducted to achieve score comparability (Kolen & Brennan, 2004).

Numerous scale linking/equating studies have been conducted for mixed-format tests by investigating various factors that might influence linking and equating results, such as equating group differences, form differences, characteristics of anchor item sets, sample sizes, and so forth (Bastari, 2000; Cao, 2008; Kim & Kolen, 2006; Kim & Lee, 2006; Kirkpartrick, 2005; Lee, He, Hagge, Wang, & Kolen, 2012; Paek & Kim, 2007; Tan, Kim, Paek, & Xiang, 2009; Tate, 2000; Walker & Kim, 2009; Wu, Huang, Huh, & Harris, 2009). However, these studies dealt with scale linking/equating issues under unidimensional item response theory (UIRT) frameworks for both dichotomous MC items and polytomous FR items.

Tate (2000) showed that the use of only MC-anchor items in equating with mixed-format tests can be valid only when the test is unidimensional. In other words, if a mixed-format test is not unidimensional, a UIRT scale linking/equating procedures can lead to invalid results. Also, if

the underlying local independence assumption of UIRT models is violated for mixed-format tests, standard UIRT scale linking/equating procedures based on that assumption might result in biased scale linking/equating relationships (Lee, Kolen, Frisbie, & Ankenmann, 2001; Li, Bolt, & Fu, 2005). Multidimensional scale linking/equating procedures may produce more accurate scale linking/equating results than unidimensional procedures if data for a test are multidimensional.

Some previous studies indicated that data for a mixed-format test were not unidimensional and that different item types tended to measure somewhat different constructs (Cao, 2008; Lee & Brossman, 2012; Li, Lissitz, & Yang, 1999; Sykes, Hou, Hanson, & Wang, 2002; Tate, 2000; Yao & Boughton, 2009). Most previous studies (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Thompson, Nering, & Davey, 1997; Yon, 2006) that have examined multidimensional item response theory (MIRT) in the context of accomplishing score comparability have been focused on scale linking, but not equating.

Few studies exist in the literature that investigated equating under a MIRT framework. Brossman and Lee (2013) may be the first study dealing with MIRT observed and true score equating procedures with a multidimensional two parameter logistic model. Recently, Lee and Brossman (2012) proposed observed-score equating procedures under a simple-structure MIRT (SS-MIRT) framework for mixed-format tests. They showed that the SS-MIRT procedure produced more accurate equating results than did the UIRT procedure.

The present study was designed to investigate MIRT number-correct score equating for mixed-format tests using a bi-factor MIRT model. Full-information item bi-factor analysis (Gibbons & Hedeker, 1992; Gibbons et al., 2007) has been increasingly applied and implemented as an important statistical method in psychological and educational measurement (Cai, Yang, & Hansen, 2011; DeMars, 2006; Reise, Morizot, & Hays, 2007; Rijmen, 2010). The full-information item bi-factor model provides a useful psychometric framework for mixed-format tests. One way to take item-format effects into account is to incorporate specific dimensions or factors (i.e., a specific factor for the MC-item format and another specific factor for the FR-item format) in addition to a general dimension (i.e., a general factor measured by both MC and FR formats). Figure 1 depicts the data structure for a test composed of mixed-format items where M and F represent MC and FR items, respectively; and θ_G, θ_M, and θ_F represent the general, MC-specific, and FR-specific factors, respectively. The model displayed in

Figure 1 is referred to here as a bi-factor multidimensional item response theory (BF-MIRT) model.

In the terminology of the factor analysis literature, a one-factor solution as a potential model for the item responses of a mixed-format test may not entirely reflect the underlying structure of the data. It is plausible that within-format residual dependence still remains even after controlling for the influence of the primary general factor. In a BF-MIRT model, the residual dependence is modeled as additional latent variables (i.e., specific factors).

Equating procedures based on the BF-MIRT framework have not been developed previously. The main purpose of the present study is to propose observed-score equating procedures for mixed-format tests under the BF-MIRT framework. In addition, we evaluate the proposed equating procedures using several empirical data sets with different levels of multidimensionality contributable to item formats. The specific objectives of this study are as follows:

(1) To develop procedures for equating scores from mixed-format tests under the BF-MIRT framework;

(2) To compare equating results from the BF-MIRT and UIRT frameworks with those from target equating equivalents to investigate the performance of the proposed method; and

(3) To examine the relationship between equating results from UIRT and BF-MIRT models and degrees of multidimensionality of mixed-format tests.

BF-MIRT Observed-Score Equating Procedure

In this section, detailed procedures for the BF-MIRT observed-score equating are presented. Also, other related, important issues are discussed.

BF-MIRT Models for a Mixed-Format Test

Because a mixed-format test contains both MC and FR items, two different IRT models are needed for dichotomously- and polytomously-scored items. A bi-factor extension of the classical two-parameter logistic UIRT model for MC items is used in this study, which is given by Equation 1 below (Cai, du Toit, & Thissen, 2012; Cai et al., 2011):

$$\Pr(y = 1|\theta_G, \theta_M) = \frac{1}{1+\exp\{-[d_j+a_{Gj}\theta_G+a_{Mj}\theta_M]\}} \quad (1)$$

where d_j = the intercept of item *j*, a_{Gj} = the slope of item *j* on general ability θ_G, and a_{Mj} = the slope of item *j* on MC-related ability θ_M.

A model for polytomously-scored FR items used in this paper is a bi-factor extension of the logistic version of Samejima's (1969) graded response model (GRLM; Samejima, 1997). The probability density function for the bi-factor GRLM is often stated with a cumulative probability function as shown in Equation 2 below (Cai et al., 2011; Cai et al., 2012):

$$\begin{cases} \Pr(y \geq 0|\theta_G, \theta_F) = 1 \\ \Pr(y \geq 1|\theta_G, \theta_F) = \frac{1}{1+\exp\{-[d_{1j}+a_{Gj}\theta_G+a_{Fj}\theta_F]\}} \\ \Pr(y \geq k|\theta_G, \theta_F) = \frac{1}{1+\exp\{-[d_{kj}+a_{Gj}\theta_G+a_{Fj}\theta_F]\}} \end{cases}, \quad (2)$$

where k = 1, 2, 3, …., K represents score categories; d_{kj} = the intercept of category *k* on item *j*; a_{Gj} = the slope of item *j* on general ability θ_G; and a_{Fj} = the slope of item *j* on FR-related ability θ_F. Then, a category response probability function can be obtained from a difference of two adjacent cumulative response probabilities, which is given by

$$\Pr(y = k|\theta_G, \theta_F) = \Pr(y \geq k|\theta_G, \theta_F) - \Pr(y \geq k+1|\theta_G, \theta_F). \quad (3)$$

Scale Linking

Item and ability parameters need to be placed on the same scale via a scale linking procedure prior to conducting IRT equating. While UIRT scale linking involves two linking coefficients to adjust for the origin and unit of measurement, MIRT scale linking typically considers a transformation matrix to adjust for (1) translation, (2) dilation, (3) correlation, and (4) rotation (Reckase, 2009; Thompson et al., 1997). The first two terms, translation and dilation, are similar to the origin and unit of measurement in UIRT, respectively. Various MIRT scale linking procedures have been developed (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Thompson et al., 1997; Yon, 2006).

Scale linking is closely related to data collection designs. The equivalent groups design and the common-item nonequivalent groups design are the ones that are most frequently used in equating and scale linking. In UIRT scale linking, randomly equivalent groups are assumed to possess the same ability distribution, and thus no scale linking is required to place item and ability parameter estimates on the same scale. Arbitrarily setting the origin and unit of parameter estimates from different test forms to the same values automatically places parameter estimates

on the same scale. In the same vein, it is not necessary to adjust for translation and dilation in MIRT scale linking with the equivalent groups design.

However, item and ability parameter estimates are still subject to rotational and correlational indeterminacy in MIRT scale linking even with the equivalent groups design (Thompson et al., 1997). By assuming a certain multivariate ability distribution, the issue regarding correlations among latent traits can be resolved. For example, by fixing multidimensional traits to be multivariate normally distributed and uncorrelated, which is denoted as MVN(**0**,**I**) with a mean vector **0** and an identity variances-covariance matrix, the correlational indeterminacy issue can be eliminated. The assumption of zero correlations among traits is consistent with the assumption of the BF-MIRT model that specifies the general and specific factors to be all mutually uncorrelated.

The rotational indeterminacy still remains unresolved for MIRT scale linking when applying the BF-MIRT model with an uncorrelated multivariate normal distribution of trait variables. However, if we implement an observed-score equating procedure using an uncorrelated trivariate normal distribution for the general and two specific MC and FR factors, the rotational indeterminacy issue can be eliminated in generating conditional distributions of number-correct scores (Brossman & Lee, in press). The marginal observed-score distribution of test scores will be the same "before" and "after" rotation of coordinates, if correlation, translation, and dilation issues are resolved.

In short, for the equivalent groups design, the BF-MIRT observed-score equating procedures can be readily conducted without any scale linking process by assuming an uncorrelated multivariate distribution for the general and specific factors. This is the same assumption used in modeling for BF-MIRT (Cai, Yang, & Hansen, 2011). Because the main purpose of this study is to develop number-correct score equating procedures under the BF-MIRT framework, we use the equivalent groups design for real data examples reported in this paper to avoid potentially complex MIRT scale linking issues in equating. However, the observed-score equating procedures proposed in this study can be applied to other equating designs such as the common-item nonequivalent groups design after taking an additional step of scale linking before equating.

True-Score Equating

Using item parameter estimates and ability distributions that are on the same scale, either IRT true- or observed-score equating procedure can be applied to find equating relationships of number-correct scores. In IRT true-score equating, true score on one form is considered to be equivalent to true score on the other form associated with a particular ability θ value. That is, two test characteristic curves of two test forms, which define the relationships between ability θ and true scores, are used to determine equating relationships of number-correct scores. The Newton-Raphson method is generally implemented to find the equivalents of number-correct scores associated with specified θ values (Kolen & Brennan, 2004).

Because MIRT models incorporate multiple latent traits (i.e., a vector of θs), it is generally necessary to conceptualize an "arbitrary" unidimensional latent trait to place test characteristic curves of two test forms on the same scale for true-score equating. Thus, when MIRT models are selected and used for modeling item responses and parameter estimation, a unidimensional approximation or reduction process can be used to find equivalents in the IRT true-score equating context.

For the case of a BF-MIRT model, the general ability might be used as an "arbitrary" reference dimension for IRT true-score equating. The influence of the other specific factors could be aggregated and averaged into the general factor to define the probability of getting a specific score given a general ability θ_G using following equation:

$$\Pr\left(X_j = x \middle| \theta_G\right) = \int \Pr\left(X_j = x \middle| \theta_G, \theta_s\right) h(\theta_s) d\theta_s, \quad (4)$$

where X_j = score on item j; θ_G= general ability; θ_s= specific ability; and $h(\theta_s)$ = probability density function for the specific ability. This equation can be applied to every item within a BF-MIRT model because all items are influenced by only two factors, the general and specific abilities. $\Pr(X_j = x|\theta_G, \theta_s)$ in Equation 4 can be determined by Equation 1 for MC items and Equation 2 for FR items, respectively. In fact, Equation 4 defines an item response function on the general factor marginalized with specific factors. Then, the conventional IRT true-score equating procedures described in Cook and Eignor (1990) and Kolen and Brennan (2004) can be implemented.

It should be noted that the aforementioned process of true-score equating under the BF-MIRT framework is based on a strong assumption that θ_G is the primary latent trait being

measured by the test and the other specific latent traits, $\theta_s's$, are considered trivial or nuisance dimensions for test developers and/or users. It might be difficult, however, to defend the use of the "arbitrary" unidimensional approximation or reduction process when applying the BF-MIRT model to modeling item responses and parameter estimation.

For this reason, IRT observed-score equating might be preferred to IRT true-score equating with the BF-MIRT approach in that the former does not require a unidimensional approximation process and uses a selected BF-MIRT model in parameter estimation and equating in a consistent way. Therefore, we focus our investigation on the performance of BF-MIRT observed-score equating in this study.

Observed-Score Equating

BF-MIRT observed-score equating can be accomplished with following three steps: (1) generate conditional observed score distributions, (2) aggregate the conditional distributions to marginal observed score distributions, and (3) find equipercentile equating relationships. Based on item parameter estimates of the BF-MIRT model, the conditional observed score distribution at each combination of θ_G (general ability), θ_M (MC-specific factor), and θ_F (FR-specific factor) from a mutually uncorrelated trivariate normal distribution can be obtained using an extended version of the recursive formulas by Lord and Wingersky (1984) for dichotomous items and Hanson (1994) for polytomous items.

A marginal observed score distribution is obtained by aggregating conditional observed score distributions over the entire trivariate theta distribution, $g(\theta_G, \theta_M, \theta_F)$, as follows:

$$f(x) = \iiint_{-\infty}^{\infty} f(x|\theta_G, \theta_M, \theta_F) g(\theta_G, \theta_M, \theta_F) d\theta_G d\theta_M d\theta_F. \quad (5)$$

The integration in Equation 5 can be approximated by summation with a specified set of quadrature points and weights as

$$f(x) = \sum_{\theta_G} \sum_{\theta_M} \sum_{\theta_F} f(x|\theta_G, \theta_M, \theta_F) q(\theta_G, \theta_M, \theta_F), \quad (6)$$

where $q(\theta_G, \theta_M, \theta_F)$ is the density function of the trivariate normal distribution. After the marginal observed score distributions for both old and new test forms are obtained, the traditional equipercentile equating method is applied to find the equating relationships.

Method

Data Source

Data for this study were from the College Board Advanced Placement (AP) examinations. Three types of data sets were formulated to evaluate the performance of the BF-MIRT and UIRT observed-score equating methods for mixed-format tests: matched samples, pseudo forms, and simulated data sets. Descriptive statistics of these three data sets are presented in Table 1, and detailed explanations about formulation of the three data sets are provided next.

As indicated in the previous section, the equivalent-group data collection design is considered in this study to avoid relatively complex issues related to the MIRT scale linking. The equating design for the operational AP exams, however, is a common-item nonequivalent groups design (CINEG), in which data are collected for two sample groups from different populations. Therefore, some arbitrary manipulations of actual data are desired to create data sets that are similar to those that would have been obtained from the equivalent groups design.

Data Set 1: Matched Samples (Pseudo Groups). Matching techniques have been used in experimental designs to reduce differences in experimental and control groups, which also decrease random errors and increase statistical power. In applying matching techniques to equating data from the CINEG design, it has been suggested that use of common-item scores be avoided. Livingston, Dorans, and Wright (1990) indicated that matching on common-item scores leaded to biased equating results. Wright and Dorans (1993) proposed using other "selection variables" (i.e., previously known variables that groups differ on) and obtained more accurate equating results. Yu, Livingston, Larkin, and Bonett (2004) applied logistic regression to assign propensity scores to examinees using variables such as gender, ethnicity, etc. Recently, Powers and Kolen (2012) investigated several matching techniques to achieve more accurate equating results and found that matching on the selection variable, with or without other variables, produced more accurate equating results compared to unmatched samples.

In this study, we used three selection variables including reduced fee, parental education, and gender to create matched samples by randomly selecting the same number of examinees within each category of these variables. Data from two forms (old and new) of the AP US History and AP Chemistry exams administered in different years were used. Pseudo groups of matched samples were created and treated as equivalent groups for equating purposes.

Data Set 2: Pseudo Forms. In creating equivalent groups from data obtained based on nonequivalent groups, the matched-sample method (i.e., pseudo groups) has an advantage of using original intact forms. However, it has a potential difficulty of achieving an exact target level of group equivalence. Unlike the matched-sample method, the pseudo-form approach involved splitting a single operational test form into two halves, which resulted in two short pseudo forms (old and new). In this study, odd- and even-numbered items were used to split the original operational form and thus the two pseudo forms had the same number of unique MC and FR items. There were no common items.

A sample of 6,000 examinees was selected randomly from a large pool of examinees who took the operational test form. Li and Lissitz (2000) showed that, for a sample of 2,000 with 20 anchor items, item parameters were adequately recovered for the test response function linking procedures. The selected sample size of 6,000 in this study was three times as large as the one suggested by Li and Lissitz (2000) so as to eliminate calibration issues due to small sample size for Data Set 2, and also Data Set 3 of simulation.

Item responses for the 6,000 selected examinees to the odd-numbered items constituted one data set for the old form, and item responses for another different 6,000 selected examinees to the even-numbered items formed the data set for the new form. Thus, these pseudo form data represented a situation where the two different groups of examinees took either old or new form. One form of each of AP Art History and AP Spanish exams was used to formulate data for pseudo forms.

Data Set 3: Simulation. It is reasonable to assume that Data Set 2 for pseudo forms is similar to those from randomly equivalent groups and provides an appealing criterion equating relationship as discussed later. However, the pseudo forms are shorter and thus do not reflect psychometric characteristics of the actual forms. For example, the pseudo forms created for Data Set 2 had a relatively small number of FR items, which might lead to underestimating format effects in the context of equating. In order to make up for the shortcomings for the matched samples and pseudo-form data and to investigate the relationship between equating results and degree of multidimensionality of mixed-format tests, simulation techniques were also considered.

A simple structure MIRT model was used for generating response data under the assumption that mixed-format tests measure two correlated factors, one for MC items and

another for FR items (Kolen, Wang, & Lee, 2012; Lee & Brossman, 2012). The rationale for using the simple structure MIRT model was to eliminate unintended advantages of using the same base simulation model in comparing several different models. For example, a major comparison made in this study is between UIRT and BF-MIRT models. If the UIRT model is used as a base simulation model, the equating results for the UIRT model will be preferred over the results for the BF-MIRT model, and vice versa. The use of the simple structure MIRT model as a base simulation model was considered to be relatively neutral compared to the use of either the UIRT or BF-MIRT model. The simple structure MIRT model is also useful for generating data responses with some degree of local dependence due to item formats and to manipulate the degree of multidimensionality by specifying different levels of correlation between MC and FR latent traits. In this study, six levels of correlation were considered, 0.50, 0.60, 0.70, 0.80, 0.90, and 0.99, ranging from a highly multidimensional case to a nearly unidimensional case.

Two forms of the AP World History exam were used to obtain item parameters for the simulation. Item parameters of 70 MC items and 3 FR items (scored 0 to 9 each) were estimated using samples of 6,136 for the new form and 5,952 for the old form. Two separate calibrations were carried out using IRTPRO (Cai, du Toit, & Thissen, 2012) to estimate separately item parameters of the two parameter logistic model for the MC items and the graded response model for the FR items. The resulting item parameter estimates were assumed to be item parameters for the simple structure model. A bivariate standard normal distribution with specified correlation was used to generate response data for 6,000 examinees for each form. Weights of 1 and 2 were assigned to MC and FR items, respectively, which made the maximum total score equal to 124 (70 score points for MC items plus 54 score points for FR items). These weights were considered to make the total test scores as similar as possible to the operational test scores.

Analyses and Evaluation

UIRT and BF-MIRT observed-score equating procedures as well as traditional equipercentile equating were considered in this study. For UIRT equating, item parameters of the two-parameter logistic model for MC items and the graded response model for FR items were estimated using IRTPRO (Cai et al., 2012) and UIRT observed-score equating relationships were found using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). For the BF-MIRT model equating, item parameters were estimated using IRTPRO (Cai et al., 2012) and the observed-score equating relationships were computed using a program written for this purpose. To

minimize effects of using different IRT estimation programs, IRTPRO was used for both UIRT and BF-MIRT models.

Target equating equivalents were determined using the traditional equipercentile equating method for matched-sample data (i.e., Data Set 1) and simulation (i.e., Data Set 3). The rationale for using the traditional equipercentile equating as a criterion was that (a) it does not favor one IRT model over the other, and (b) it would not produce biased results due to multidimensionality of the data, at least in theory. For pseudo-form data (i.e., Data Set 2), target equating equivalents were established by the traditional equipercentile equating method using a single group design for the entire group of examinees who took the operational test. The large-sample single-group equipercentile equating for the second data set provided an appealing criterion in that it was based on a very large sample size, there was no group difference, and equating relationships were determined by directly linking two full-length forms as opposed to linking through a common-item set. Different target equating equivalents were used for three types of data to compare the performances of the UIRT and BF-MIRT observed-score equating methods.

The overall level of discrepancy between target equating equivalents and the UIRT and BF-MIRT equating results was computed using two indices. The first weighted root mean squared difference index denoted here as WRMSD1 was computed by Equation 7 below:

$$\text{WRMSD1} = \{\sum_1^K w_i \, (EEQ_i - TEQ_i)^2\}^{\frac{1}{2}}, \tag{7}$$

where EEQ_i is the estimated equating equivalent of raw score *i* on the new form based on a particular equating method (i.e., either UIRT or BF-MIRT); TEQ_i is the target equating equivalent of raw score *i* on the new form; *K* represents the maximum score; *i* represents each raw score point; and w_i is the weight of frequency of raw score *i* in the new-form data.

The second index was the weighted root mean squared difference (WRMSD2) and is given by

$$\text{WRMSD2} = \{\sum_1^K w_i \, (BFEEQ_i - UIRTEQ_i)^2\}^{\frac{1}{2}}, \tag{8}$$

where $BFEEQ_i$ is the estimated equating equivalent of raw score of *i* on the new form based on BF-MIRT observed-score equating; $UIRTEQ_i$ is the estimated equating equivalent of raw score

of *i* on the new form using the UIRT observed-score equating; and all other variables are the same as defined in Equation 7.

Model Data Fit Statistics

Several model data fit statistics were used to provide additional comparative information between UIRT and BF-MIRT models. Because the item-parameter estimation method implemented in this study was marginal maximum likelihood, -2 log-likelihood (-2LL) measure and related Akaike's Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC; Schwarz, 1978) were compared. The smaller values of these fit statistics, the better the model fits data. The -2LL and AIC tend to favor more complex models with large sample sizes. As expected, in all cases of this study, -2LL and AIC for the BF-MIRT model were smaller than those for the UIRT model, and will not be discussed any further here. The BIC gives some penalty for more complex models with large samples (DeMars, 2012) and is used in this study for comparing model data fits of the UIRT and BF-MIRT models.

The BF-MIRT model captures residual dependence after controlling for the influence of the primary factor. If residual dependence exists, it indicates the violation of local independence assumption when applying the UIRT model to the data. Although the local independence assumption cannot be directly tested because the latent trait cannot be observed, many statistics have been proposed to provide information about local dependence among items and/or subsets of items (Chen & Thissen, 1997; Maydeu-Olivarves & Joe, 2005, 2006; Yen, 1984). In this study, the M_2 statistics with standardized bivariate residuals based upon a full marginal table was used for comparing the UIRT and BF-MIRT models in terms of model-data fit (Liu & Maydeu-Olivares, in press).

Results

Equating Results for UIRT and BF-MIRT Procedures

Figure 2 shows equating results of the UIRT and BF-MIRT methods using the traditional equipercentile method as a baseline with Data Set 1: Matched Samples. The horizontal axis represents raw scores with a score range truncated to scores with more than 0.5 percent of examinees at each tail. The vertical axis represents the differences between the equivalents of either the UIRT or BF-MIRT method and the equivalents of the baseline method. The straight zero line on the vertical axis represents results for the criterion equipercentile equating method. For the AP Art History exam, the equivalents from the BF-MIRT equating were closer to the

criterion results in the score range from 40 to 70. The equivalents for the UIRT and BF-MIRT methods for the AP Chemistry exam were almost indistinguishable.

Equating results for the UIRT and BF-MIRT methods for Data Set 2: Pseudo Forms are displayed in Figure 3. Again, results for the two IRT methods along with the criterion equipercentile equating are presented in the figure. The results for the UIRT and BF-MIRT methods for the AP Art History exam were almost identical. By contrast, relatively distinct differences between the UIRT and BF-MIRT methods were found for the AP Spanish exam, in which a closer relationship of the BF-MIRT method to the criterion equipercentile equating were observed in the score range from 25 to 35 compared to the UIRT method.

Similar analyses were conducted with Data Set 3: Simulation using different levels of correlation. Figure 3 shows equating results of applying the UIRT and BF-MIRT methods to six simulated data sets. The case of correlation equal to 0.99 might be viewed as the simulated data responses being likely to be unidimensional. On the contrary, it would be reasonable to expect that data become more multidimensional as the correlation between MC and FR latent traits decreases. Closer relationships of equating results between the UIRT and BF-MIRT methods were found for conditions with higher correlation values. In particular, it was difficult to differentiate equating equivalents between the UIRT and BF-MIRT methods for the case of correlation equal to 0.99. For the conditions with correlation values between 0.50 and 0.80, the BF-MIRT method provided closer equating equivalents to those of the criterion equipercentile equating method, except for correlation of 0.70. By contrast, the UIRT method tended to provide better results than the BF-MIRT method for correlation values of 0.70 and 0.90.

Equating Results and Degree of Multidimensionality

The two evaluation criteria, WRMSD1 and WRMSD2, for the UIRT and BF-MIRT methods are presented in Table 2 for Data Sets 1, 2, and 3. WRMSD1 shows an overall discrepancy between either the UIRT or BF-MIRT method and the criterion equipercentile equating method. WRMSD2 represents an overall difference between the UIRT and BF-MIRT methods. The WRMSD1 for the UIRT and BF-MIRT methods were 0.75401 and 0.72548, respectively, for the AP US History of Data Set 1. This means that the equivalents of the BF-MIRT method were closer to those of the criterion equipercentile method than the UIRT method. Similar values of WRMSD1 between the UIRT and BF-MIRT methods were reported for the AP Chemistry of Data Set 1. The WRMSD2 for the US History and Chemistry exams were 0.06366

and 0.03258, respectively, which indicates that the differences between the UIRT and BF-MIRT methods were larger for US History than Chemistry. Very similar patterns were found with the AP Spanish and the AP Art History exams for Data Set 2.

The simulation analysis using Data Set 3 was intended to investigate the relationship between equating results and degree of multidimensionality of mixed-format tests. For the simulation condition with correlation of 0.50, it is reasonable to assume some non-negligible departure from unidimensionality in the data due to format effects. For this condition, the smaller WRMSD1 of the BF-MIRT method was reported compared to the UIRT method. Similar results were found for other correlation values up to 0.80, except for a correlation of 0.70. For correlation values of 0.70 and 0.90, the WRMSD1 of the UIRT method was smaller than that of the BF-MIRT method. Both UIRT and BF-MIRT methods provided very similar WRMSD1 values for correlation of 0.99. In general, smaller WRMSD1 of the BF-MIRT method was observed for lower correlation conditions. Given the fact that the criterion equipercentile equating method is prone to standard error of equating to some extent, it might not be so surprising to see irregular trends for some cases such as correlation of 0.70. Although more replications would have led to more stable results, the results reported in this paper seemed sufficient to demonstrate the relative performances of the two IRT methods. Recall that the main purposes of this study were to develop the BF-MIRT observed-score equating procedure for mixed-format tests and to investigate relative appropriateness of the proposed method. Because calibration with the BF-MIRT model on one data set requires a fair amount of time (often over an hour), a more extensive simulation study including various conditions was left for future research.

More consistent results were observed for the WRMSD2 index in relation to correlation between MC and FR latent traits. The WRMSD2 value tended to decrease as the correlation value increased. This finding is graphically presented in Figure 5. It is clear from Figure 5 that the correlation between MC and FR latent traits and overall weighted discrepancies between the UIRT and BF-MIRT methods show a constant decreasing pattern.

The differences between the UIRT and BF-MIRT equating methods conditional on raw-score points across the six correlation conditions are plotted in Figure 6. The zero line indicates the results for the UIRT method. Obviously, the conditional differences between the two methods become smaller and smaller as the correlation value gets higher.

Model Data Fit Statistics for UIRT and BF-MIRT Models

Table 3 presents Bayesian information criterion (BIC) and M_2 model-data-fit statistics for the UIRT and BF-MIRT models for the old and new forms of Data Sets 1, 2, and 3. In evaluating model data fit, a model with the lower values of BIC or M_2 would be preferred. Using both -2LL and AIC indices (not reported here), the BF-MIRT model provided better model data fit in all cases compared to the UIRT model. As previously mentioned in the Method section, these two indices tend to prefer more complex models with larger sample sizes.

With the BIC index, the BF-MIRT model provided better model data fit than did the UIRT model, except for the old form of AP Art History and the old and new forms of simulated data with correlation of 0.90 and 0.99. Relatively large differences in M_2 values between the UIRT and BF-MIRT models were reported for the old form of AP US History, old and new forms of AP Spanish. Differences of the M_2 statistics for the simulated data sets were greater than 1,000 up to correlation of 0.80, and about 450 and 180 for correlation of 0.90 and 0.99, respectively.

The results of the model data fit statistics seemed consistently related to the equating results of the UIRT and BF-MIRT models. For example, for data with relatively poor fit of the UIRT model and relatively good fit of the BF-MIRT model such as AP US History, Spanish, and simulation data with correlations up to 0.80, the BF-MIRT method provided equating results that were closer to the criteria. For other cases of relatively small differences of model data fit statistics for the UIRT and BF-MIRT models, both UIRT and BF-MIRT equating methods produced very similar equating results.

Conclusions

Within-format residual dependence may remain after controlling for the influence of the general factor for the mixed-format tests. Item-format effects can be taken into account by introducing specific MC and FR factors in addition to a general factor in the BF-MIRT model. This study was aimed to develop a BF-MIRT model observed-score equating procedure for mixed-format tests and to evaluate its performance relative to the UIRT procedure using various types of mixed-format tests. Based upon results of this study, the following conclusions can be made.

Either a true- or observed-score equating procedure can be implemented in IRT equating to develop equating relationships of number-correct scores. To determine MIRT true-score

equating relationships, it would be necessary to apply an "arbitrary" unidimensional approximation or reduction process to obtain test characteristic curves that have a one-to-one relationship between true scores and thetas. MIRT observed-score equating, by contrast, does not require an arbitrary unidimensional approximation or reduction process and thus might be preferred over true-score equating. In addition, without using a unidimensional approximation, the observed-score equating uses the full MIRT model in the process of modeling data responses, parameter estimation, and equating in a consistent way.

In many cases considered in this study, the BF-MIRT observed-score equating method produced equating results that were similar to those for the UIRT observed-score equating method. Consequently, the differences between the two methods might be viewed as practically insignificant, in general. This conclusion can be interpreted as indicative of very minor within-format residual dependence for mixed-format tests after controlling for the influence of the primary general factor. Either the UIRT or BF-MIRT observed-score equating method can be considered to find equating relationships for mixed-format tests in most practical situations. However, this result should not be over-generalized to other contexts where other factors such as testlets and contents constitute sub-dimensions.

In investigation of the relationship between equating results and degree of multidimensionality of mixed-format tests, the UIRT and BF-MIRT methods provided somewhat different equating results in cases of relatively low correlation (i.e., less than 0.80) between MC and FR traits. The BF-MIRT equating method was found to produce better equating results for mixed-format tests when a certain degree of multidimensionality exists. The better performance of the BF-MIRT method can be supported by model data fit analyses. That is, the BF-MIRT method tends to provide better equating results compared to the UIRT method when the test data are fit relatively better by the BF-MIRT model than the UIRT model.

The scope of this study was limited to the use of IRT observed-score equating and examples were considered with the random groups design. In order to implement the BF-MIRT equating procedure under a common item non-equivalent groups (CINEG) design, a scaling linking process needs to be conducted prior to equating. It is important to note that successful scale linking with a BF-MIRT model under the CINEG design requires a common-item set that consists of both MC and FR item types to properly transform all item parameter estimates for both item types. When data for a mixed-format test are obtained under the CINEG design *and* a

common-item set consists of both item types, a multidimensional scale linking procedure (see Oshima et al., 2009) can be applied to place parameter estimates for a BF-MIRT model on the same MIRT ability scale, and then the BF-MIRT equating procedure discussed in this paper can be conducted. However, it is often the case that a mixed-format test involves a common-item set that is composed solely of MC items due to many practical constraints.

Future research should investigate application of the BF-MIRT equating method to other bi-factor situations. For example, a BF-MIRT model can be fitted to tests composed of testlets, tests with several content areas, and language tests involving several sub-skills. Another interesting topic would be to compare several relatively simple MIRT models such as simple structure MIRT, BF-MIRT, and higher-order MIRT models with compensatory or non-compensatory specifications in the context of various psychometric analyses including equating.

References

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale*. Unpublished doctoral dissertation, University of Massachusetts.

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.

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*, 460-481.

Cai, L., du Toit, S. H. C., & Thissen, D. (2012). *IRTPRO*: *Flexible professional item response theory modeling for patient reported outcomes* [Computer software]. Chicago: SSI International.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221–248.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets*. Unpublished doctoral dissertation, University of Maryland.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373-399.

Davey, T., Oshima, T., & Lee, T. (1996). Linking multidimensional item calibration. *Applied Psychological Measurement, 20*, 405-416.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,* 145–168.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items*. Unpublished research note.

Hirsch, T. (1989). Multidimensional equating. *Journal of Educational Measurement, 26,* 337-349.

Kim, S., & Kolen, M. J. (2006). Robustness for format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*, 357-381.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*, 53-76.

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating*. Unpublished doctoral dissertation, University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*, 1-20.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen & W. Lee (Eds.) *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, *25*, 357-372.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied psychological measurement*, *29*(5), 340-356.

Li, Y., & Lissitz, R. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115-138.

Liu, Y., & Maydeu-Olivares, A. (in press). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*.

Livingston, S. A., Dorans, N. J., & Wrights, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73-95.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 452-461.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables. *Journal of the American Statistical Association, 100,* 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713-732.

Min, K. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations*. Unpublished doctoral dissertation, Michigan State University.

Oshima, T., Davey, T., & Lee, K. (2009). Multidimensional linking: four practical approaches. *Journal of Educational Measurement, 37*, 357-373.

Paek, I., & Kim, S. (2007). *Empirical investigation of alternatives for assessing scoring consistency on constructed response items in mixed format tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Powers, S., & Kolen, M. J. (2012). Using matched samples equating methods to improve equating accuracy. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361-372.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). *Multidimensionality and the equating of a mixed-format math examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Tan, X., Kim, S., Paek, I., & Xiang, B. (2009). *An alternative to the trend scoring shifts in mixed-format tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37,* 329-346.

Thompson, T., Nering, M., & Davey, T. (1997). *Multidimensional IRT scale linking*. Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Walker, M., & Kim, S. (2009). *Linking mixed-format tests using multiple choice anchors*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating.* (ETS Research Report RR-93-4). Princeton, NJ: Educational Testing Service.

Wu, N., Huang, C., Huh, N., & Harris, D. (2009). *Robustness in using multiple-choice items as an external anchor for constructed-response test equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*, 177-197.

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

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling.* Unpublished doctoral dissertation, Michigan State University.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essay.* (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.

Table 1

Descriptive Statistics of Three Data Sets Used in This Study

Test	Form	No. of MC Items	No. of FR Items	Max Score Point	No. of Examinees	Mean	SD
			Data Set 1: Matched Samples				
US History	new	79	3 (9,9,9)	106	9,245	54.5	16.27
	old	79	3(9,9,9)	106	9,245	53.6	16.31
Chemistry	new	74	4(10,10,9,8)	111	3,240	55.3	22.63
	old	74	4(10,10,9,8)	111	3,240	57.9	20.46
			Data Set 2: Pseudo Forms				
Art History	new	55	4(9,4,4,4)	76	6,000	42.0	11.67
	old	55	4(9,4,4,4)	76	6,000	45.0	10.18
Spanish	new	34	2(5,5)	44	6,000	29.7	6.73
	old	34	2(5,5)	44	6,000	29.0	6.53
			Data Set 3: Simulation				
Original	new	70	3(9,9,9)	124	6,136	62.3	22.64
	old	70	3(9,9,9)	124	5,952	61.0	21.43
R=0.50	new	70	3(9,9,9)	124	6,000	62.4	20.54
	old	70	3(9,9,9)	124	6,000	61.7	19.15
R=0.60	new	70	3(9,9,9)	124	6,000	62.6	20.83
	old	70	3(9,9,9)	124	6,000	60.7	19.75
R=0.70	new	70	3(9,9,9)	124	6,000	62.5	21.50
	old	70	3(9,9,9)	124	6,000	60.8	20.09
R=0.80	new	70	3(9,9,9)	124	6,000	62.5	22.07
	old	70	3(9,9,9)	124	6,000	61.3	20.69
R=0.90	new	70	3(9,9,9)	124	6,000	62.4	22.36
	old	70	3(9,9,9)	124	6,000	61.0	21.31
R=0.99	new	70	3(9,9,9)	124	6,000	62.4	23.32
	old	70	3(9,9,9)	124	6,000	60.9	21.78

Table 2

WRMSD1 and WRMSD2 of UIRT and BF-MIRT Equating Methods for Data Sets 1, 2, and 3

Test	**WRMSD1**		**WRMSD2**
	UIRT	BF-MIRT	
	Data Set 1: Matched Samples		
US History	0.75401	0.72548	0.06366
Chemistry	1.41168	1.41511	0.03258
	Data Set 2: Pseudo Forms		
Art History	0.15138	0.15344	0.00297
Spanish	0.12408	0.11931	0.03521
	Data Set 3: Simulation		
R=0.50	0.44368	0.32654	0.17134
R=0.60	0.24271	0.22944	0.16249
R=0.70	0.09289	0.13924	0.13451
R=0.80	0.26684	0.24326	0.09250
R=0.90	0.22946	0.25496	0.06134
R=0.99	0.22192	0.22805	0.01452

Note. WRMSD = weighted root mean squared difference; UIRT = unidimensional item response theory; BF-MIRT = Bi-factor multidimensional item response theory.

Table 3

Bayesian Information Criterion (BIC) and M_2 Model-Data-Fit Statistics

		BIC		M_2		
Test	Form	UIRT	BF-MIRT	UIRT	BF-MIRT	Difference
		Data Set 1: Matched Samples				
US History	New	917438	916397	11589.7	9496.0	2093.7
	Old	937560	927252	31924.3	8643.1	23281.2
Chemistry	New	324166	323814	10357.0	8730.2	1626.8
	Old	312653	312532	10059.1	9065.4	993.7
		Data Set 2: Pseudo Forms				
Art History	New	447201	447146	5204.8	4519.8	685.0
	Old	411989	412122	4642.4	4239.7	402.7
Spanish	New	245376	242865	7625.2	2748.0	4877.2
	Old	255720	253920	6054.5	2634.7	3419.8
		Data Set 3: Simulation				
R=0.50	New	533546	530699	7115.7	4088.7	3027.0
	Old	532242	529990	6899.4	4239.6	2659.8
R=0.60	New	533921	531719	6803.0	4278.3	2524.7
	Old	533851	531913	6324.8	4145.6	2179.2
R=0.70	New	531977	530485	5928.2	3995.2	1933.0
	Old	532845	531707	5772.7	4039.3	1733.4
R=0.80	New	529697	528983	5530.4	4354.2	1176.2
	Old	529967	529345	5198.8	4052.4	1146.4
R=0.90	New	528175	528286	4549.9	4094.4	455.5
	Old	527027	527180	4582.7	4148.1	434.6
R=0.99	New	523739	524233	4338.5	4160.6	177.9
	Old	524941	525429	4348.1	4165.6	182.5

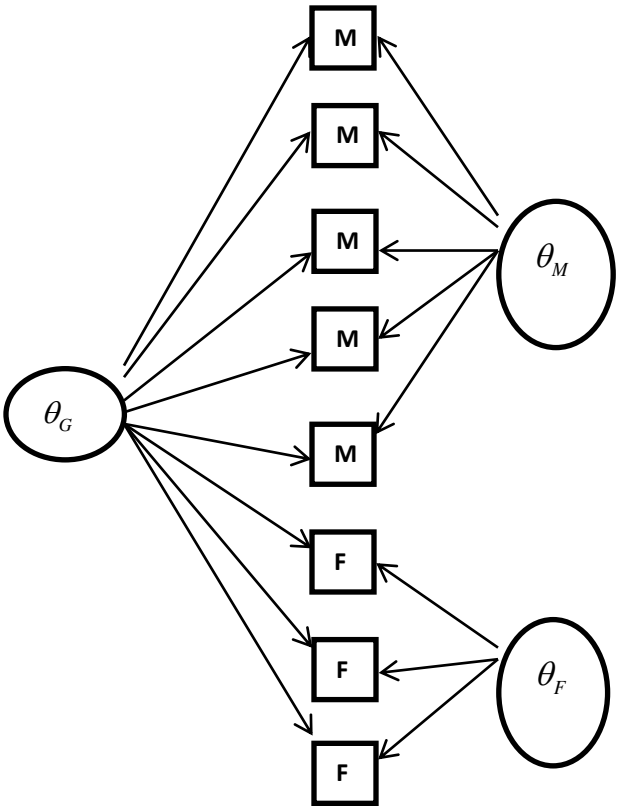
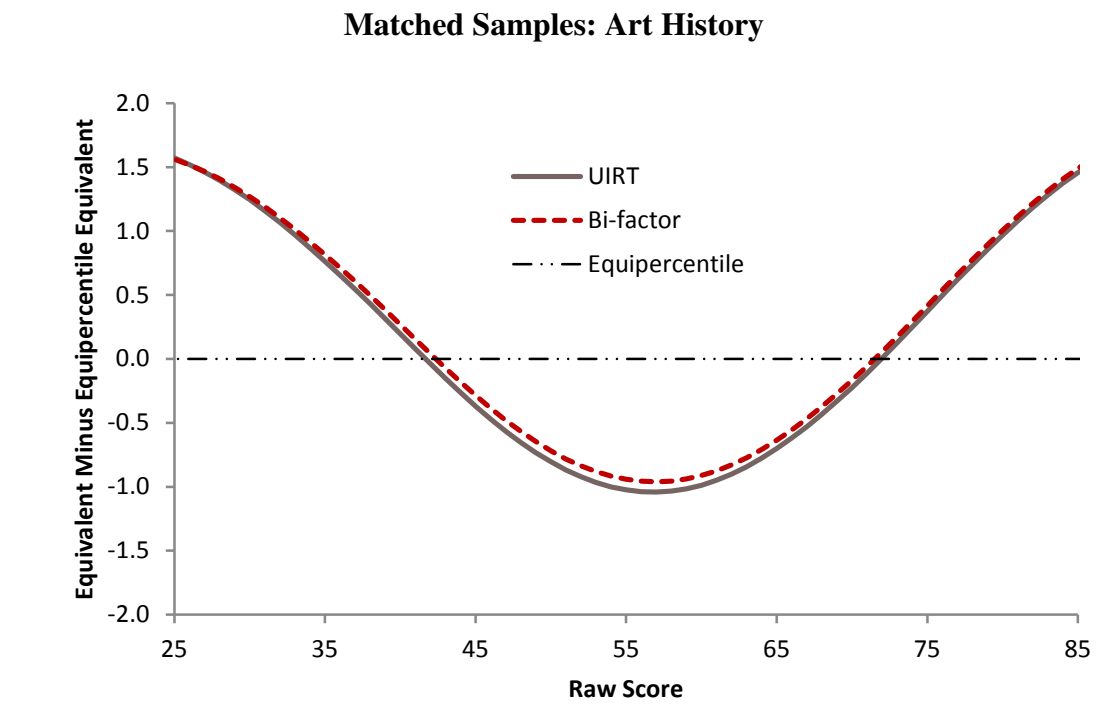


Figure 1. Bi-factor model for a mixed-format test.



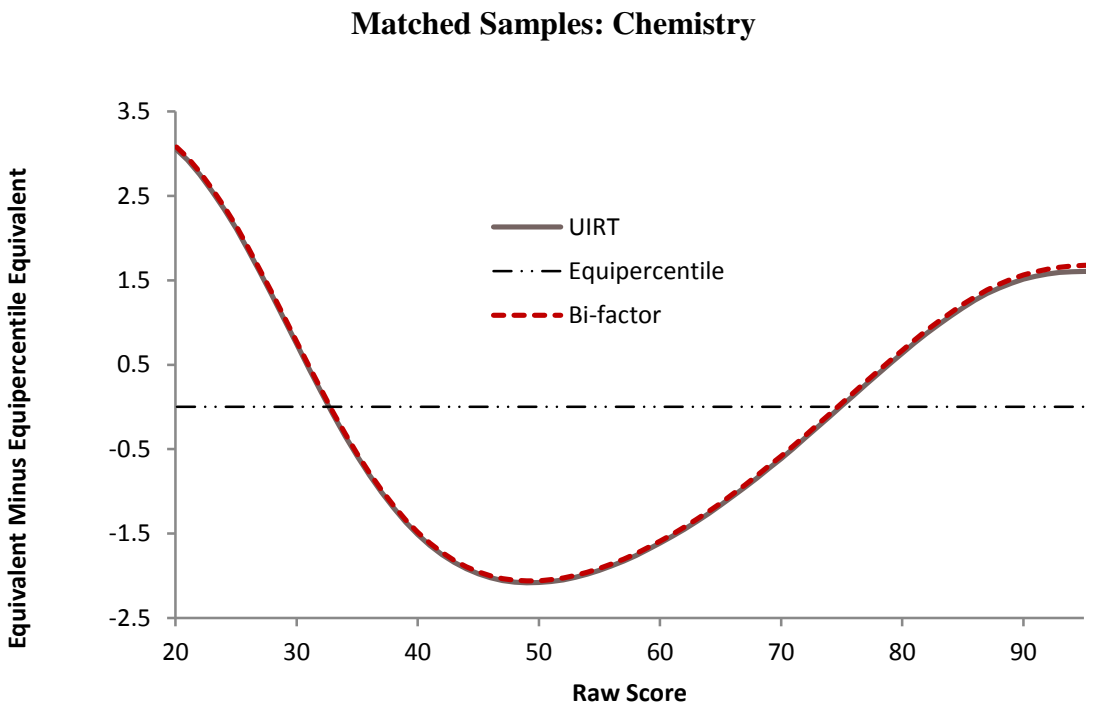


Figure 2. Equating results of UIRT and BF-MIRT methods using equipercentile method as a baseline with Data Set 1: Matched Samples.

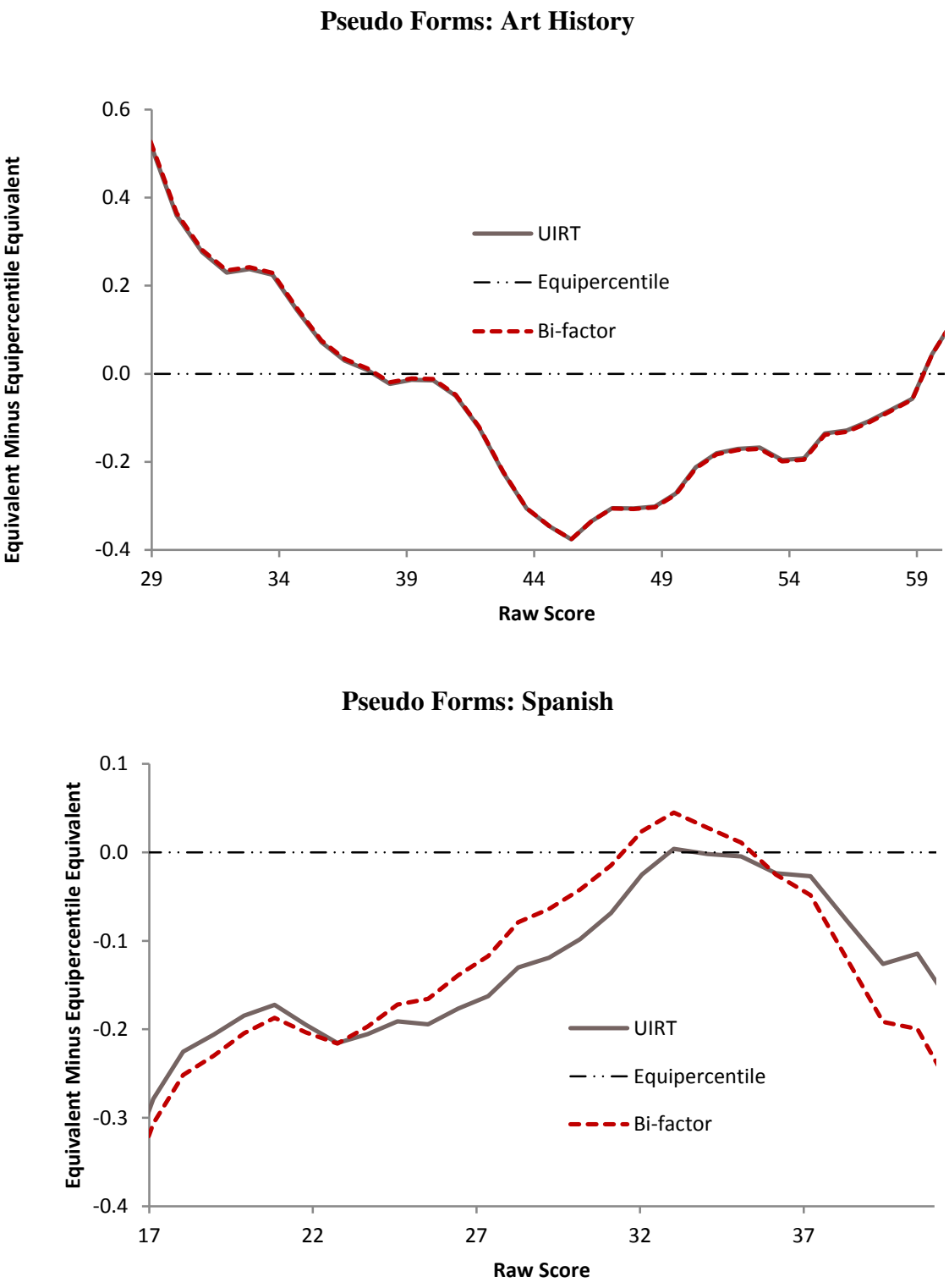


Figure 3. Equating results of UIRT and BF-MIRT methods using equipercentile method as a baseline with Data Set 2: Pseudo Forms.

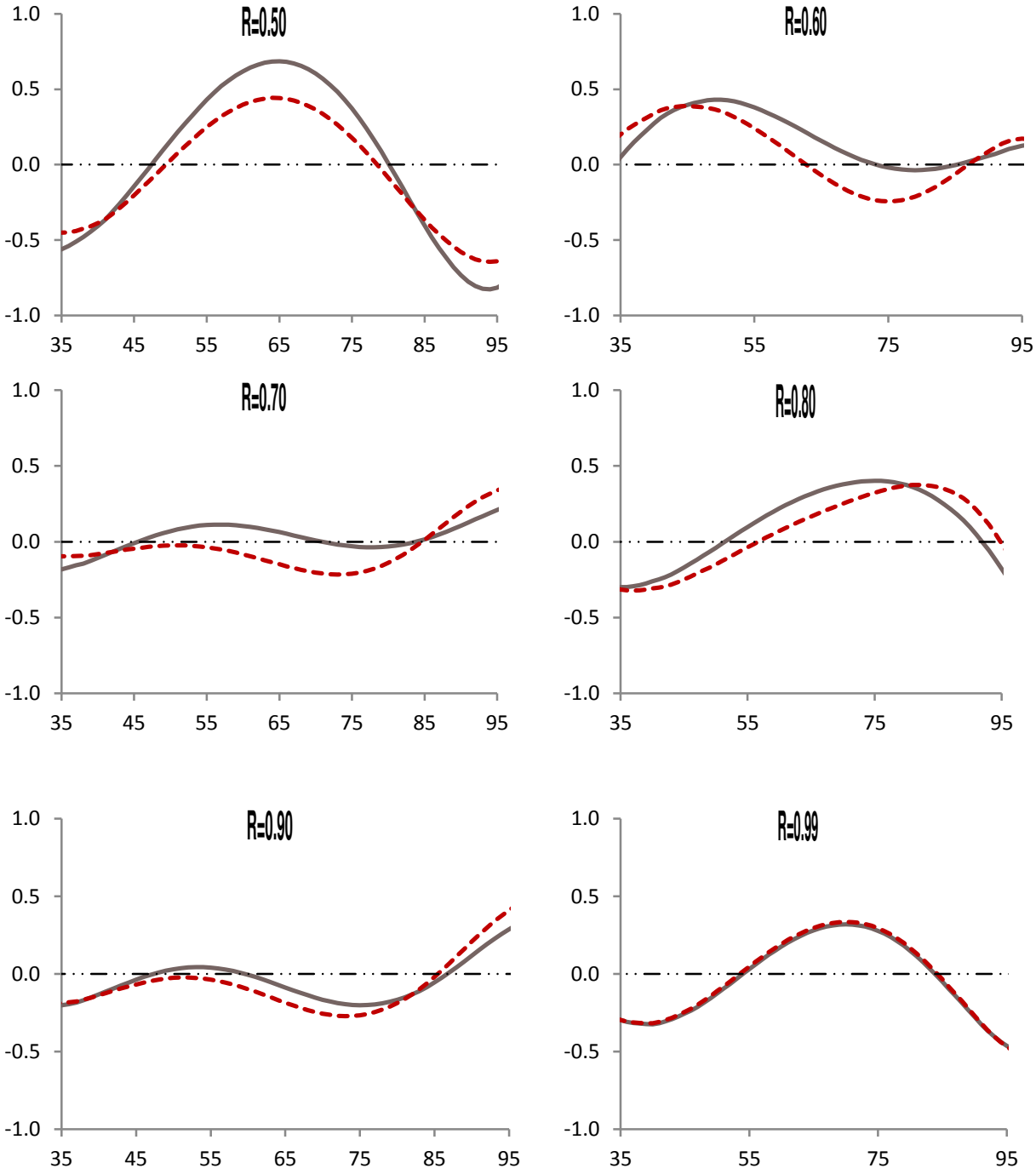


Figure 4. Equating results of UIRT and BF-MIRT methods using equipercentile method as a baseline with Data Set 3: Simulation.

Note. Horizontal axis = raw score; vertical axis = difference between equivalents of either UIRT or BF-MIRT and equipercentile equivalents; solid line = UIRT; dotted line = BF-MIRT; zero line = criterion equipercentile.

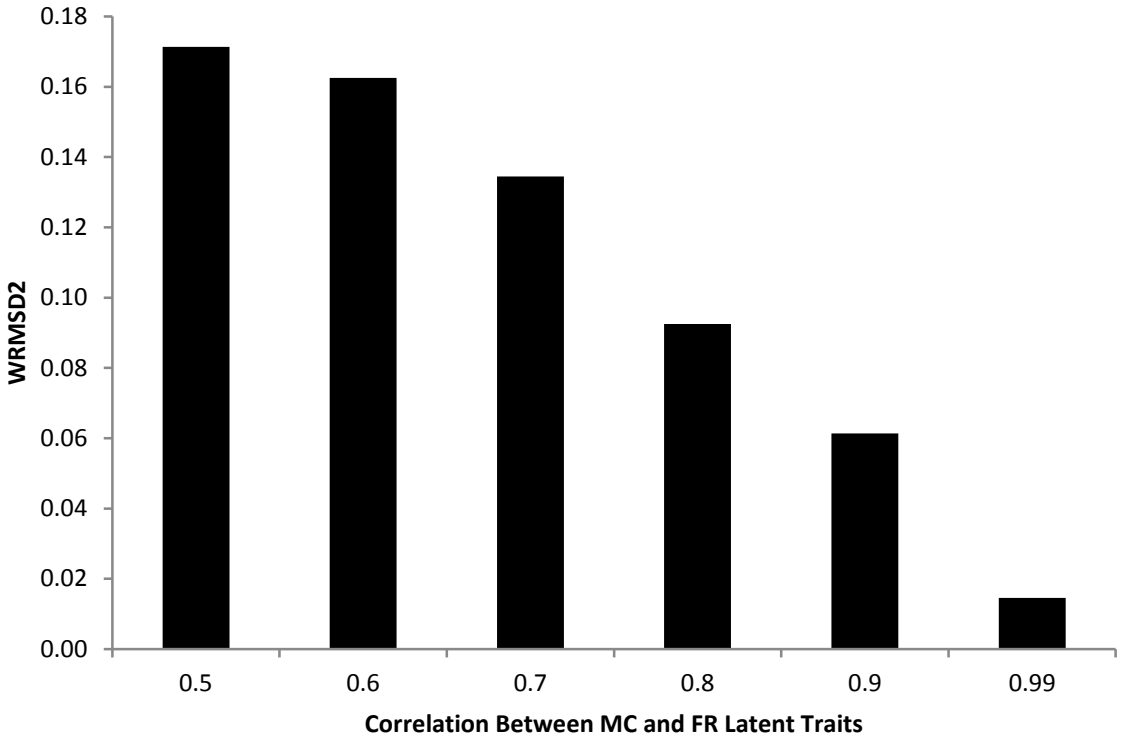


Figure 5. Correlation between MC and FR latent traits and corresponding weighted root mean squared difference index 2 (WRMSD2).

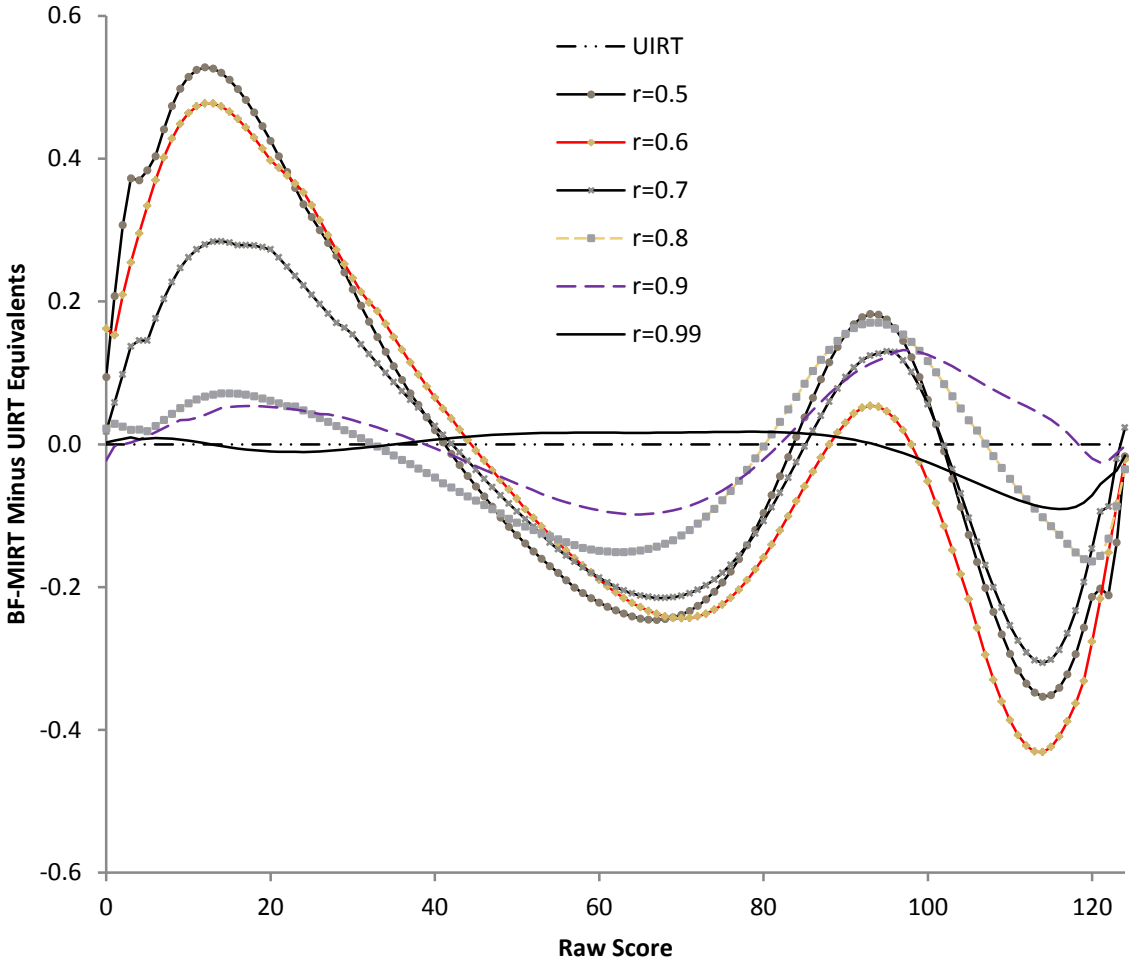


Figure 6. Differences between UIRT and BF-MIRT equating methods on raw score point for specified correlation between MC and FR latent traits.

Chapter 8: Multidimensional Item Response Theory Observed Score Equating Methods for Mixed-Format Tests

Jaime Peterson and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

The purpose of this study was to build upon the existing MIRT equating literature by introducing a full multidimensional item response theory (MIRT) observed score equating method for mixed-format exams. At this time, the MIRT equating literature is limited to full MIRT observed score equating methods for multiple-choice only exams and Bifactor observed score equating methods for mixed-format exams. Given the high frequency with which mixed-format exams are used and the accumulating evidence that some tests are not purely unidimensional, it was important to present a full MIRT equating method for mixed-format tests.

The performance of the full MIRT observed score method was compared with the traditional equipercentile method, and unidimensional IRT (UIRT) observed score method, and Bifactor observed score method. With the Bifactor methods, group-specific factors were defined according to item format or content subdomain. With the full MIRT methods, two- and four-dimensional models were included and correlations between latent abilities were freely estimated or set to zero. All equating procedures were carried out using end-of-course exams in Spanish Language and English Language and Composition. For each subject two separate datasets were created using pseudo-groups in order to have two separate equating criteria. The specific equating criteria that served as baselines for comparisons with all other methods were the theoretical Identity equating line and the traditional equipercentile procedure.

In general, the multidimensional methods were found to perform better for datasets that evidenced more multidimensionality, whereas unidimensional methods worked better for unidimensional datasets. In addition, the scale on which scores are reported influenced the comparative conclusions made among the studied methods. For performance classifications, which are most important to examinees, there typically were not large discrepancies among the UIRT, Bifactor, and full MIRT methods. However, this study was limited by its sole reliance on real data which was not very multidimensional and for which the true equating relationship was not known. Therefore, plans for improvements, including the addition of a simulation study to introduce a variety of dimensional data structures, are also discussed.

Multidimensional Item Response Theory Observed Score Equating Methods for Mixed-Format Tests

For many testing programs it is often necessary to administer alternate forms of a test due to circumstances such as multiple testing dates and threats to test security. It is desirable that those forms can be given interchangeably across situations, which often requires the use of equating methods. The use of IRT equating is particularly appealing in situations where pre-equating is necessary.

Even though the use of IRT equating methods can provide greater flexibility in comparison to traditional methods, they still have limitations. For example, the majority of IRT methods assume item-level unidimensionality, meaning that each item measures one construct. In order for IRT equating methods to be useful, the assumptions made by these methods need to hold. In general, UIRT equating results have been found to be robust to moderate (but not severe) violations of the unidimensionality assumption (Bolt, 1999; Camilli, Wang, & Fesq, 1995; Cook et al., 1985; de Champlain, 1996; Dorans & Kingston, 1985; Yen, 1984). Even with this finding, equating should be conducted in such a manner so that the effects of violations of assumptions are minimized (Kolen & Brennan, 2014). In situations where unidimensionality does not hold, alternate methods need to be available. If the situation requires pre-equating, then IRT methods would still be desirable but a multidimensional IRT (MIRT) model may be more appropriate.

The use of a MIRT model instead of UIRT model is one manner in which assumption violations can be lessened if evidence suggests the data are not unidimensional. It is likely that MIRT models provide a better fit than UIRT models in some instances (Li, Li, & Wang, 2010), and as a result, MIRT scale linking and equating procedures are becoming more prevalent. However, there is still a lot of research that needs to be done, especially in the context of mixed-format exams. Therefore, the main objective of this study is to expand upon the MIRT observed score equating literature in the context of mixed-format tests. Specially, the main objectives are to: (1) present observed score equating methods for mixed-format tests using a full MIRT modeling framework; (2) compare the differences in equating results from using full MIRT, Bifactor MIRT, UIRT, and traditional equipercentile methods; (3) compare differences in equating results for the full MIRT method when the correlations between latent traits are specified as zero versus when they are freely estimated; and (4) examine whether dimensionality

due to item format versus dimensionality due to content subdomain influences the relationships among the four equating methods in a similar manner.

Theoretical Framework

Item Response Theory

Item response theory (IRT) is a useful psychometric tool because it allows for predictions to be made about how examinees will answer test items without actually having to administer them (Lord, 1980). Using IRT, the probability that an examinee with a certain ability level will correctly answer an item with particular characteristics can be expressed with an item response function (IRF).

Unidimensional models. The IRF for the two-parameter logistic (2PL) and three-parameter logistic (3PL) models (Birnbaum, 1968) respectively, can be expressed as

$$P_i(\theta_j) = \frac{1}{1 + e^{-1.7a_i(\theta_j - b_i)}} \tag{1a}$$

and

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_j - b_i)}}. \tag{1b}$$

Here, a_i represents the discrimination of item i, b_i is the item's location, c_i is the item's pseudo-guessing parameter, θ_j is the latent ability of person j, and 1.7 is a scaling constant (i.e., D) used to make results comparable to the normal ogive metric. The 2PL and 1PL models are special cases of the 3PL model and all are used with dichotomous or multiple-choice (MC) test items.

IRT models have also been developed for free-response (FR) items that have more than two score categories. Polytomously-scored items have the benefit of providing more information about an examinee's ability level in comparison to dichotomously scored items, but at the cost of requiring greater testing time. This gain in information, along with other considerations such as improved score validity, has led many state and national testing programs to use mixed-format tests that consist of both MC and FR item types. Two of the most commonly used polytomous IRT models are Samejima's (1969) graded response (GR) model and Muraki's (1992) generalized partial credit (GPC) model.

The GR model (Samejima, 1969) was developed for items with ordered response categories. To compute the probability that an examinee will respond in one of K categories for item i, first the cumulative category response function must be calculated such as,

$$P_{ik}^{*}(\theta_j) = 1, \qquad k = 1,$$
$$P_{ik}^{*}(\theta_j) = \frac{e^{[Da_i(\theta_j - b_{ik})]}}{1 + e^{[Da_i(\theta_j - b_{ik})]}}, \qquad k = 2, \ldots, K_i. \tag{2a}$$

Here, b_{ik} is the location of the k^{th} category boundary for item i, and all other parameters can be interpreted consistently using definitions previously provided. The cumulative response function is interpreted as the probability that an examinee will score in category k or higher. In order to compute the probability that an examinee will respond in a particular category, the difference between cumulative categories is computed as,

$$P_{ik}(\theta_j) = P_{ik}^{*}(\theta_j) - P_{i(k+1)}^{*}(\theta_j), \qquad k = 1, \ldots, K_i - 1,$$
$$P_{ik}(\theta_j) = P_{ik}^{*}(\theta_j), \qquad k = K_i. \tag{2b}$$

Multidimensional Item Response Theory

MIRT is an extension of UIRT and was developed to more accurately estimate the relationship between items and examinees in situations where items measure more than one latent trait or different traits. The need for MIRT stemmed from findings that UIRT models can oversimplify, and thereby misrepresent, the relationship between items and examinees when the latent test space is multidimensional. Furthermore, several studies have suggested that different item formats tend to measure different constructs (Lee & Brossman, 2012; Li, Lissitz, & Yang, 1999; Tate, 2000; Yao & Boughton, 2009); however this difference is generally ignored when UIRT models are used. With MIRT models, individual items can measure several abilities at once (i.e., complex structure) or measure one of several abilities represented by the test (i.e., simple structure). The MIRT modeling framework likely resembles the underlying dimensional structure of tests more accurately, especially mixed-format tests.

Similar to UIRT models, MIRT models make assumptions concerning functional form, local independence, and dimensionality (de Ayala, 2009). In the MIRT framework, the local independence assumption is conditional on a vector of latent abilities rather than just one.

MIRT models. MIRT models can be classified as either compensatory or noncompensatory depending on whether the number of latent traits (m) are specified to have a multiplicative or additive relationship as modeled by the IRF. For noncompensatory models, the relationship is multiplicative such that levels of the m latent traits both impact the response probability. For compensatory models, a high level of one trait can make up for, or compensate for, deficiencies on another trait because the relationship is additive. The current study focuses only on compensatory MIRT models wherein a linear combination of θ coordinates is used to

predict response probabilities. This decision was made based on software availability and the fact that compensatory MIRT models are more prevalent in the literature.

The multidimensional 3PL (M3PL; Reckase, 1985) model is an extension of the UIRT 3PL model and includes a pseudo-guessing parameter which results in a non-zero lower asymptote. Here, the UIRT θ is replaced with $\boldsymbol{\theta}$, which is a 1 x m vector of examinee coordinates in m-dimensional space; and a is replaced with $\mathbf{a}$, which is a 1 x m vector of discrimination parameters. The parameter d is a scalar and represents the item's intercept. The IRF for the M3PL is expressed as

$$P_i\left(U_i = 1 \middle| \boldsymbol{\theta_j}, \boldsymbol{a_i}, c_i, d_i\right) = c_i + (1 - c_i)\frac{e^{\boldsymbol{a_i}\boldsymbol{\theta}_j' + d_i}}{1 + e^{\boldsymbol{a_i}\boldsymbol{\theta}_j' + d_i}}. \tag{3}$$

The M3PL model can be made comparable to the normal ogive metric by multiplying the term, $\mathbf{a_i}\boldsymbol{\theta_j} + d_i$, in the numerator and denominator by the scaling constant $D = 1.7$.

MIRT models also exist for items with more than two response categories and are referred to as polytomous MIRT models, and are extensions of their UIRT counterparts. Muraki and Carlson (1993) developed a multidimensional extension of the graded response (MGR) model that is expressed in the normal ogive metric. Similar to the UIRT GR model, the MGR model assumes ordered response categories and that successful completion of step k requires that step k –1 be successfully completed first. The minimum and maximum scores for item i are zero and K_i, respectively. The probability of scoring at each of the k steps is essentially modeled by the M2PL model with responses below k scored as 0 and those equal to or greater than k scored as 1. It should be noted that the M2PL model is a special case of the M3PL model where $c_i = 0$. Then, the probability of scoring in a particular category, k, is found by subtracting the probability of scoring at the $k + 1$ step from the probability of scoring at step k, as given by

$$P_i\left(U_i = k \middle| \boldsymbol{\theta_j}\right) = P_i^*\left(U_i = k \middle| \boldsymbol{\theta_j}\right) - P_i^*\left(U_i = k + 1 \middle| \boldsymbol{\theta_j}\right). \tag{4}$$

The cumulative category response probability for $U_i = 0$ is always 1, and for $U_i = K_\mathrm{i} + 1$ is always 0. The mathematical form of the normal ogive MGR model is given as

$$P_i\left(U_i = k \middle| \boldsymbol{\theta}_\mathbf{j}\right) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{a_i'\theta_j} + d_{i,k+1}}^{\mathbf{a_i'\theta_j} + d_{ik}} e^{\frac{-t^2}{2}}\, \mathrm{d}t. \tag{5}$$

Here, k represents the examinee's score on item i and can take the values of 0, 1, . . ., K_i; $\mathbf{a_i}$ is a vector of discrimination parameters; and d_ik is a parameter representing the easiness of reaching

the k^{th} step of the item. The d_{ik} parameter for a score of zero is set to $-\infty$ and for a score of $K_i + 1$ is set to ∞.

The Bifactor model was first introduced by Gibbons and Hedeker (1992) and differs from the aforementioned MIRT models in that items are allowed to load on one general dimension and one group-specific dimension, and is confirmatory in nature. An example of a factor pattern for a three-item test under the Bifactor model is

$$\begin{bmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & 0 & a_{32} \end{bmatrix}.$$

It can be seen that all items load on the general or first dimension, while the first two items have an additional loading on the first group-specific dimension and the third item on the second group-specific dimension. In this example, the general dimension could represent the subject area of science while the specific dimensions represent item format such that the first two are MC items and the third is an FR item. This modeling approach allows for the residual variance due to item formats to be taken into account.

Under the Bifactor MIRT model, the assumption of conditional independence is made with respect to all latent dimensions, general and group-specific. The probability of a correct response given an examinee's general (θ_0) and specific (θ_S) latent abilities under the Bifactor extension of the UIRT 3PL model was provided by Cai, Yang, and Hansen (2011) as,

$$P_i(U_i = 1|\theta_0, \theta_S) = c_i + \frac{1 - c_i}{1 + e^{\{-D[d_i + a_{i0}\theta_0 + a_{iS}\theta_S]\}}}. \tag{6}$$

Here, d_i represents the item intercept, a_{i0} represents the slope of the item on the general dimension, a_{iS} is the item slope for the group-specific dimension, and D is a scaling constant.

Cai et al. (2011) also presented a logistic version of the Bifactor extension of the GR model that is similar to the MGR model provided by Muraki and Carlson (1993). For an item with K response categories, let $U_i \in \{0, 1, \ldots, K-1\}$ so that the cumulative response probability is

$$P_i^*(U_i \geq K - 1|\theta_0, \theta_S) = \frac{1}{1 + e^{\{-D[d_{i,K-1} + a_{i0}\theta_0 + a_{iS}\theta_S]\}}}. \tag{7}$$

There is a total of $K - 1$ cumulative response functions for an item with K response categories. The notation used to describe the parameters in the model is the same as previously defined. As with the UIRT and MIRT GR models, the probability of responding *in* a particular category

under the Bifactor GR model is found by taking the difference between two adjacent cumulative response functions.

For mixed-format tests, the conditional category response probability for examinee *j*'s response to item *i* under the Bifactor model is $P_i(U_i = k|\theta_0, \theta_S)$ for $k = 0, \ldots, K_i - 1$ and takes on a multinomial distribution. Cai et al. (2011) provide the conditional density function for an item as

$$f(x|\theta_0, \theta_S) = \prod_{k=0}^{K_i-1} P_i(U_i = k|\theta_0, \theta_S)^{\chi_k(x_{ij})}. \quad (8)$$

The indicator function, $\chi_k(x_{ij})$, takes on the value 1 when $U_i = k$ and 0 otherwise, thereby pulling out the category response probabilities corresponding to examinee *j*'s observed responses.

Dimensionality Assessments

In order to determine whether a UIRT or MIRT equating method is more appropriate, it is necessary to conduct a series of dimensionality assessments. An important point made by Reckase (2009) is that dimensionality is not a property of the test, but depends on both the sensitivity of the test items to the targeted latent traits and the variability in the examinee sample on those traits.

After having reviewed the dimensionality assessment literature, Reise et al. (2000) suggested that it is better to overestimate the dimensionality of a data structure than to underestimate it. On the other hand, as dimensionality increases so does the number of parameters that need to be estimated which in turn increases the chance for estimation error. Regardless, dimensionality assessment is a critical precursor to accurate (M)IRT equating.

(M)IRT Observed Score Equating Methods

The first step in IRT observed score equating is to estimate number-correct score distributions on both forms using an IRT model or a combination of models in the case of mixed-format tests (Kolen & Brennan, 2014). The IRT model can be unidimensional or multidimensional, depending on the data. In the case of the former, conditional distributions of observed number-correct scores for a given θ_j are modeled using the compound binomial and multinomial distributions for dichotomous and polytomous models, respectively. Traditional equipercentile methods are then used to equate the two model-based estimated observed score distributions (see Kolen & Brennan, 2014 for more details).

MIRT observed score equating is conducted in a similar manner as UIRT observed score equating except that observed score distributions are conditional on a vector of latent abilities rather than a single ability. Furthermore, the ability density used to obtain the marginal observed score distributions is multivariate instead of univariate (Brossman & Lee, 2013).

Several studies have presented MIRT equating methods in recent years. For example, Lee and Brossman (2012) presented observed score equating procedures under a simple-structure (SS) MIRT framework for mixed-format tests and compared its results with those obtained using UIRT and traditional equipercentile methods. In their study, dimensionality was associated with item format such that θ_1 and θ_2 represented MC and FR items, respectively. The steps involved were much like UIRT observed score equating, with the exceptions of (a) MC and FR item types were calibrated separately, (b) a bivariate latent ability distribution was specified, and (c) the conditional observed score distribution for each item type had to be estimated separately and then combined across item types. Results from real data analyses indicated that between the UIRT and SS-MIRT procedures, the latter was found to produce results most similar to the traditional equipercentile method, and led to less total equating error when used with multidimensional data.

The first full MIRT equating procedure was illustrated by Brossman and Lee (2013), but for MC-only tests. Specifically, they provided examples of three equating methods: (1) Full MIRT observed score equating, (2) unidimensional approximation to MIRT observed score equating, and (3) unidimensional approximation to MIRT true score equating. Both UIRT approximation methods were extensions of earlier work by Zhang and colleagues (Zhang, 1996; Zhang & Stout, 1999; Zhang & Wang, 1998). In general, they found that multidimensional procedures performed more similarly to traditional equipercentile methods, while the UIRT method resulted in more systematic error. Interestingly, both unidimensional approximation methods produced equating results that were very similar to the Full MIRT method.

Recently, a study by Lee and Lee (2014) presented a MIRT observed score equating procedure for mixed-format tests using the Bifactor model. Again, multidimensionality was associated with item format. The 2PL and GR Bifactor models were fit to a series of mixed-format tests using three types of datasets: (1) matched samples from intact groups, (2) pseudo-forms from a single group, and (3) a simulation study. An orthogonal trivariate distribution for each combination of the general and group-specific traits was specified by using an extended

version of the Lord and Wingersky (1984) and Hanson (1994) recursive formulas for dichotomous and polytomous items, respectively, and marginal observed score distributions were computed across the trivariate distribution. In general, they concluded that when the Bifactor model fit the data better, equating results favored the Bifactor method over the UIRT method.

After reviewing the literature, there is a recognizable need for further research on equating multidimensional tests. While there have been a handful of MIRT equating studies, none have presented solutions for conducting full MIRT equating with mixed-format tests. Furthermore, the added complexity of incorporating non simple-structure MIRT models and MIRT models for polytomous items is likely to be a topic of interest for many testing programs. Therefore, the purpose of this study was to address these complexities and provide illustrations of full MIRT observed-score equating procedures for mixed-format tests.

Method

Data and Procedures

The data used in this study were from the College Board's Advanced Placement (AP) Exams administered in 2011, and were originally collected under the common-item nonequivalent groups (CINEG) design. Specifically, the main and alternate forms for Spanish Language and English Language and Composition were chosen so that different content areas would be represented. Each exam is mixed-format and contains MC items, all with five response options, and FR items with various numbers of response categories.

Even though AP Exams were used in this study, adjustments were made that changed the psychometric properties of the exams in such a way that results cannot be generalized to the operational AP Exams. Specifics related to these modifications are described in more detail later. Furthermore, the conversion tables used in this study were devised for research purposes only, and do not reflect the operational conversion tables. As a result, the focus of this study is on the equating methods and not the characteristics of specific AP Exams.

The Spanish Language main and alternate form were given to a total of 118,176 and 3,243 examinees, respectively, and each contained 70 MC items and 4 FR items. For Spanish Language, 34 of the MC items were listening tasks and 36 were reading comprehension tasks and 2 FR items were writing and the other 2 were speaking. The English main and alternate forms were administered to 404,970 and 5,141 examinees, respectively, and each consisted of 54 MC items and 3 FR items. The possible score points and operational weights for the Spanish

Language and English items on the main and alternate forms, can be found in Table 1. Even though the alternate and main forms shared a set of common items, they were treated as operational items in order to maintain test length and adhere to the operational test specifications.

Number-correct scoring was carried out in such a way that missing and incorrect responses received a score of zero. Instead of using the operational non-integer weights, integer weights were rounded to the nearest whole number. As a result, the weights and maximum score points used in this study do not match those used operationally (Table 1).

Operationally, weights are applied to examinees' MC and FR section scores to obtain composite raw scores, which are then transformed to AP grades. In this study, composite raw scores were transformed to composite scale scores that ranged from 0 to 70, and to AP Grades that ranged from 1 to 5. The scale scores were developed to be normally distributed with a mean of 35 and standard deviation of 10, for the main form. Once the alternate form (i.e., new form) composite raw scores were equated to the main form (i.e., old form) scale, equated composite scores were transformed to scale scores using a raw-to-scale conversion table. Similarly, equated composite raw scores from the alternate form were converted to AP Grades using conversion tables developed for the main form.

Matched Samples Datasets

Given that the AP Exams were collected under the CINEG design and this study focused on the random groups (RG) design, data manipulation was required in order to attain randomly equivalent groups. Therefore, a subset of examinees was sampled from the main and alternate form groups using a selection variable (i.e., ethnicity with seven levels) that was related to students' total scores. Sampling was done with the intent to create two matched subsamples that were similar enough in ability to be considered randomly equivalent.

In order to evaluate the extent to which randomly equivalent groups were actually attained, measures of effect size for common item scores were computed between alternate and main form subgroups. The formula used to compute the effect size for group differences was a modification of the Cohen's *d* statistic that uses a pooled standard deviation. The target effect size in this study was specified as ± 0.05 standard deviations.

Since real data were used in this study, the population equating relationship could not be known. Therefore, an estimate of the population equating relationship was used to obtain target equivalents that were used to evaluate the studied equating methods. For the Matched Samples

datasets, the population equating relationship was approximated by conducting traditional equipercentile equating with presmoothing (Lee & Lee, 2014). Using the equipercentile method to approximate the population equating relationship has been used in previous studies (Brossman & Lee, 2013; Lee & Lee, 2014), and appears to offer a theoretically sensible, yet not perfect solution. The equipercentile method is a preferable option because, unlike (M)IRT methods, it does not assume a particular dimensionality of the data. Thus, there is no theoretical reason to suspect that equipercentile results will align more closely with any of the proposed (M)IRT methods, thus reducing potential bias. Last, the equipercentile method is similar to the IRT methods in that it specifies a curvilinear equating relationship.

Single Population Datasets

Since the use of the traditional equipercentile equating method as a criterion was not without limitations, another criterion was included to corroborate results. Using the same original AP datasets, modifications were made so that each original Main Form was equated to itself. To do this, examinees were randomly sampled from the Main Form group and assigned to pseudo-Old and pseudo-New Form groups, each with sample sizes of 3,000. For these datasets, equating was theoretically unnecessary since the same form (i.e., the original Old Form) was administered to both pseudo-groups and those pseudo-groups come from the same population, which is why they were referred to as "Single Population" datasets.

For these datasets, the Identity equating line was used as the criterion where a score on the New Form is equal to that same score on the Old Form. If the New Form score were subtracted from the Old Form equivalent to create a difference graph, the Identity line would be a straight zero-difference line. This approach is not perfect however because some sampling error is expected in the estimated equating relationships. Therefore, slight departures from the Identity line were expected since samples were used instead of populations and equating procedures were not expected to adjust for sampling error.

This approach is also known as "equating a test to itself" (Kolen & Brennan, 2014) and has been used as a criterion in other equating studies when the true equating relationship was unknown (Petersen et al., 1982). However, this approach is also not without limitations and has been found to favor equating procedures with the fewest number of parameters (Brennan & Kolen, 1987a; 1987b). As a result, Kolen and Brennan (2014) recommend that this approach be used to detect poor performing equating methods rather than pinpoint the best performing one.

Traditional equipercentile with presmoothing was included as a condition with the Single Population datasets in order to determine how much it deviated from the Identity relationship. If there was a large discrepancy, this could signal that it may not be a good proxy to the population equating relationships for the Matched Sample datasets and is more sensitive to random error.

Dimensionality Assessments

Several techniques were used to thoroughly assess the dimensionality of the AP Exams in regard to item format and content subdomain, each being considered separately. Analyses were conducted separately on the main and alternate forms using each procedure discussed next. First, the disattenuated correlation between section scores (where section was defined according to item format or content specification, but not both) was computed as,

$$\rho_{T_1T_2} = \frac{\rho_{S_1S_2}}{\sqrt{(\rho_{S_1S_1'})(\rho_{S_2S_2'})}}\ , \tag{9}$$

where $\rho_{S_1S_2}$ represents the observed correlation between scores on sections 1 and 2, $\rho_{S_1S_1'}$ is the coefficient alpha reliability of section 1 scores, and $\rho_{S_2S_2'}$ is the coefficient alpha reliability of section 2 scores. When considering dimensionality associated with item format, sections 1 and 2 correspond to total scores for MC and FR items, respectively. Disattenuated correlations close to 1 indicate the two sections measure similar construct(s). If all pairwise disattenuated correlations are near 1, this can be taken as support that unidimensionality holds for the form.

Principal component analyses (PCA) on tetrachoric and polychoric correlations were also conducted to evaluate the dimensionality on the main and alternate form datasets. Correlation matrices were computed in SAS (SAS Institute, Version 9.3), while the PCAs were performed using Mplus software (Muthén & Muthén, 1998-2012). Scree plots of eigenvalues were used to determine the amount of aggregate variance accounted for by each PC or latent dimension (Cattell, 1966). Next, a series of non-linear exploratory factor analyses (EFAs) were conducted to compare solutions with various numbers of factors retained. Both, orthogonal and oblique rotations were performed to be consistent with the latent trait distributions specified in some of the equating procedures. Two sets of confirmatory factor analyses were run on each form where dimensions represented either content subdomain or item format, as outlined by the test specification table. For all analyses, the Akaike Information Criteria (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) were used to assess model fit. The AIC

and BIC can be used to compare the fit of non-nested models, and for both indices, smaller values indicate better fit.

Last, the software program, Poly-DIMTEST (Li & Stout, 1995) was used to assess whether the datasets were essentially unidimensional. The test employed by the Poly-DIMTEST program is based on estimating covariances between pairs of items, conditional on a specific subscore. The Poly-DIMTEST test statistic, adjusted for bias, was used to test the null hypothesis of essential unidimensionality.

Item Calibration and Model Data Fit

Unidimensional IRT, Bifactor, and full MIRT versions of the 3PL + GR model combination were calibrated for all AP forms using *flexMIRT* (Cai, 2012). The Bifactor model was estimated once with group-specific factors defined by item format or content subdomain. The full MIRT models were calibrated using two- and four-dimensional solutions with correlations between factors fixed at zero or freely estimated. A summary of the studied conditions can be found in Table 2. It should be noted that calibrations were repeated twice – first for the Single Population and then for the Matched Samples datasets, therefore the number of conditions is twice of that reported in Table 2.

For the unidimensional calibrations, scale indeterminacy was handled by specifying latent ability distributions as univariate standard normal (i.e., $\theta \sim UVN(0, 1)$).
For multidimensional calibrations, latent ability distributions for the New and Old Forms were specified in two different ways. First, the latent ability distributions were specified as multivariate normal with means of 0 and standard deviations of 1 for each of the m latent abilities (i.e., $\boldsymbol{\theta} \sim MVN(0, \mathbf{I})$). In this setting the correlation between latent abilities was assumed to be zero. Using the second method, distributions were specified to follow a multivariate normal distribution, but the correlations between dimensions were freely estimated during item calibration. Since forms were calibrated separately, it was possible for the estimated correlations to vary across forms. Performing a rotation of the axes would not provide an exact solution to the correlational differences because while the correlation would change, the value to which it changed could not be controlled. Furthermore, rotation changes the mean, standard deviation, and correlations between the latent dimensions, but not the coefficient in the IRF (Reckase, 2009), thereby leaving the marginal observed score distributions unchanged. If correlational differences were found, the estimated correlations could be either left unchanged or averages

could be taken. In this study, correlations were allowed to differ between groups rather than taking the average across groups. When latent traits are assumed orthogonal, only one quadrature distribution is required for equating (as opposed to two when correlations were freely estimated), because all correlations are fixed at zero.

For the majority of the methods, 25 evenly spaced quadrature points were used during estimation; however, it was necessary to reduce the number of quadrature points to 11 for the BF-content method. This reduction was made only for the estimation of the observed total score distributions, which was conducted after item parameter estimation was completed. Thus, across all models, 25 quadrature points were used during item parameter estimation.

Complex structure, as opposed to simple structure, was specified for all full MIRT conditions such that items were free to load on all *m* dimensions. For the Bifactor model, items were free to load on the general dimension and one additional group-specific dimension, resulting in two non-zero discrimination parameter estimates per item. To avoid convergence problems, prior distributions were used to estimate the slope and pseudo-guessing parameters. Prior distributions for the slope and pseudo-guessing parameters were set as lognormal (0, 0.5) and beta (5, 17), respectively, in the UIRT, full MIRT, and Bifactor calibrations. These prior settings are employed in BILOG-MG (Zimowski et al., 2003) and have been found to keep parameter estimates within plausible ranges for UIRT models. Calibration settings can be found in Table 3 and were selected to be as similar as possible across UIRT and MIRT models in order to reduce random noise.

Several fit indices in addition to results from the dimensionality assessments were used to evaluate the goodness-of-fit of the studied (M)IRT models to the AP data. The eigenvalues produced from the PCAs were first examined to judge whether the unidimensionality assumption of the UIRT models was supported. Specifically, Reckase's (2009) recommendation that in order for unidimensionality to be plausible, the first eigenvalue should account for at least 20% of the total variance, was used. A non-significant test statistic provided by Poly-DIMTEST was also taken as support that essential unidimensionality held.

Equating Procedures

As mentioned previously, Identity equating and traditional equipercentile equating with log-linear presmoothing served as proxies for the population equating relationships for the Single Population and Matched Samples datasets, respectively. In order to determine the optimal degree

of smoothing, plots of raw-to-raw equivalents were examined as well as moments for raw scores and unrounded scale scores. The polynomial degree of smoothing will be provided in the Results section. Traditional Equipercentile procedures were carried out in *RAGE-RGEQUATE* (Kolen, 2005).

UIRT observed score equating. In order to perform UIRT observed score equating, number-correct score distributions were estimated for the Old and New Form examinees using a combination of UIRT models (depending on the number of response options), which were then equated using traditional equipercentile methods. A recursive algorithm was used to compute the conditional observed score distributions which were then aggregated to form the marginal distribution. For dichotomous and polytomous items, conditional number-correct score distributions were found using the Lord and Wingersky (1984) and Hanson (1994) and Thissen et al. (1995) recursive formulas, respectively. The conditional observed score distributions are then combined to create marginal distributions, which are then equated using the traditional equipercentile method.

The unidimensional IRT observed score equating procedures were conducted in *RAGE-RGEQUATE*, using estimated marginal total score distributions from *flexMIRT*. More specifically, the marginal observed score distributions were inputted as frequencies in the *RAGE-RGEQUATE* program, along with the observed raw composite score and raw-to-scale conversion table for the Old Form. Traditional equipercentile equating was carried out in the usual manner and without smoothing.

Bifactor observed score equating. The steps involved in Bifactor observed score equating are very similar to those performed in UIRT observed score equating. For example, conditional number-correct scores were estimated and then aggregated to form marginal score distributions, which were then equated using traditional equipercentile methods. However, instead of having number-correct scores conditioned on a single θ, they were conditioned on θ_G (general ability) and group-specific θ_S abilities. The Bifactor observed-score equating procedures were also conducted by using the estimated marginal distributions from *flexMIRT* as input to *RAGE-RGEQUATE*.

Full MIRT observed score equating. A full MIRT observed score equating procedure for dichotomous items was originally presented by Brossman and Lee (2013), which was applied to MC items, and will be discussed in greater detail. The methodology required to conduct full

MIRT observed score equating including polytomous items is presented for the first time here. During item calibration and estimation of the marginal observed score distributions, the latent ability distribution was specified as either (a) multivariate normal (i.e., $\boldsymbol{\theta} \sim MVN(0, \mathbf{I})$) with fixed zero correlations, or (b) multivariate normal with correlations equal to the estimated values found during item calibration. The translation and dilation indeterminacies were handled in *flexMIRT* by fixing each coordinate axis to have a mean of 0 and standard deviation of 1. No transformations were performed to handle rotational indeterminacies under the full MIRT conditions because doing so would have no effect on response probabilities (Brossman & Lee, 2013; Reckase, 2009). Estimated correlations for each group were used to create two separate quadrature distributions (one for each group), that in turn were used to estimate marginal observed score distributions. The estimated marginal distributions were then used to equate the New Form scores to the Old Form scale, using traditional equipercentile methods.

The conditional observed score distribution on the MC section is found for each combination of m abilities and is expressed as $f_r(x \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a vector of m abilities and subscript r is the item index. The probability of correctly answering the first item is $f_1(x = 1|\boldsymbol{\theta}_j) = P_1$, whereas the probability of an incorrect answer is $f_1(x = 0|\boldsymbol{\theta}_j) = (1 - P_1)$. For a test with $r > 1$, the conditional number-correct distributions are computed using a direct extension of the Lord-Wingersky algorithm (Brossman & Lee, 2013) as

$$\begin{aligned} f_r(x|\boldsymbol{\theta}_j) &= f_{r-1}(x|\boldsymbol{\theta}_j)(1 - P_r), && x = 0 \\ &= f_{r-1}(x|\boldsymbol{\theta}_j)(1 - P_r) + f_{r-1}(x - 1|\boldsymbol{\theta}_j)P_r, && 0 < x < r, \\ &= f_{r-1}(x - 1|\boldsymbol{\theta}_j)P_r, && x = r. \end{aligned} \tag{10}$$

Equation 10 differs from its UIRT counterpart in that $\boldsymbol{\theta}_j$ has replaced θ_j. The recursion formula continues with $r - 1$ iterations for a test of length r. To obtain the marginal distributions for each form, the conditional distributions are multiplied by the multivariate ability density ($\psi(\boldsymbol{\theta})$) and integrated over all combinations of the m latent abilities as,

$$f(x) = \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_m} f(x|\boldsymbol{\theta})\psi(\boldsymbol{\theta})\, d(\boldsymbol{\theta}). \tag{11}$$

If a discrete multivariate ability distribution is assumed, integrals can be replaced with summation so that Equation 11 can be replaced with Equation 12, and the marginal distributions are obtained by,

$$f(x) = \sum_{\theta_1} \sum_{\theta_2} \dots \sum_{\theta_m} f(x|\boldsymbol{\theta})\psi(\boldsymbol{\theta}). \tag{12}$$

For polytomous items, the recursion formulas presented by Hanson (1994) and Thissen et al. (1995) can be extended to the multidimensional framework to estimate the number-correct distributions, conditional on *m* latent abilities. For a particular item *i* with *K* response categories, the probability of examinee *j* earning a score in the k^{th} category is conditional on a vector of *m* abilities ($\boldsymbol{\theta}_{\text{j}}$) and written as,

$$f_r(U_i = W_{rk}|\boldsymbol{\theta}_{\text{j}}) = P_{ik}(\boldsymbol{\theta}_{\text{j}}), \tag{13}$$

where W_{rk} is the scoring function of the k^{th} category for the r^{th} item. When $r > 1$, the probability of a score of *x* after administration of the r^{th} item is,

$$f_r(x|\boldsymbol{\theta}_{\text{j}}) = \sum_{k=1}^{K_j} f_{r-1}(x - W_{rk})\, P_{ik}(\boldsymbol{\theta}_{\text{j}}) \qquad for \;\; min_r < x < max_r. \tag{14}$$

Here, min_r and max_r represent the minimum and maximum scores possible after addition of the r^{th} item. The marginal observed score distributions are found by multiplying the conditional distributions by the multivariate ability density ($\psi(\boldsymbol{\theta})$) and integrating or summing over all combinations of the *m* latent abilities as,

$$f(x) = \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_m} f(x|\boldsymbol{\theta})\psi(\boldsymbol{\theta})\, d(\boldsymbol{\theta}) \tag{15}$$

or

$$f(x) = \sum_{\theta_1} \sum_{\theta_2} \dots \sum_{\theta_m} f(x|\boldsymbol{\theta})\psi(\boldsymbol{\theta}). \tag{16}$$

Since the AP exams are mixed-format, the conditional observed-score distributions found by use of dichotomous and polytomous recursive formulas need to be summed, conditional on $\boldsymbol{\theta}_{\text{j}}$, prior to estimating the marginal distributions.

Full MIRT equating was implemented by inputting the *flexMIRT* estimated marginal distributions as frequencies in the *RAGE-RGEQUATE* program, along with the corresponding raw scores and raw-to-scale conversion table for the Old Form. Traditional equipercentile equating was carried out in the usual manner and without smoothing.

Evaluation Criteria

Several criteria were used to evaluate the performance of the observed score equating procedures investigated in this study. First, weighted Root Mean Squared Differences

(wRMSDs) were computed between the studied equivalents and the criterion equivalent. The wRMSD was used to evaluate the aggregated discrepancies between the criterion equivalents and (M)IRT equivalents and is expressed as,

$$wRMSD = \left\{ \sum_{j=0}^{Z} w_j \left[eq_{Y_{IRT}}(x_j) - eq_{Y_C}(x_j) \right]^2 \right\}^{1/2}. \quad (17)$$

Here, $eq_{Y_{IRT}}(x_j)$ represents the equated equivalent for the raw score *j* using one of the (M)IRT procedures, $eq_{Y_C}(x_j)$ represents the criterion equivalent of raw score *j*, w_j is the empirical weight associated with that raw score *j* on the New Form, and *Z* is the maximum raw composite score.

Next, discrepancies between the studied equivalents and criterion equivalents were evaluated against a "Difference That Matters" (DTM; Dorans, Holland, Thayer, & Tateneni, 2003) criterion. According to Dorans et al. (2003), a DTM is marked by a difference of half a reported score unit. In the current study, differences between the (M)IRT procedures and the criterion procedures were evaluated using the 0.5 criterion and graphically displayed across all score points. This analysis was performed for raw scores and unrounded scale scores. In addition, the weighted average differences for each (M)IRT equating method was evaluated, with weights proportional to the empirical frequencies of the New Form group.

AP grade agreements. Another way to think about differences that matter is to ask the question: What matters most to the examinee? For AP examinees, the answer would most certainly be their rounded AP Grades. Therefore, agreement statistics were computed to determine whether the differences observed as a result of using different (M)IRT equating methods carried practical implications. Exact agreement percentages (EAPs) were computed for individual AP Grades (i.e., 1, 2, 3, 4 and 5), and then summed over to get the overall EAPs. More specifically, the formula used to compute overall EPAs was,

$$Overall\ EAP = \frac{s_{11} + s_{22} + s_{33} + s_{44} + s_{55}}{S}, \quad (18)$$

where, s_{aa} is the number of New Form raw score points that converted to a grade of *a* (where *a* can be 1, 2, 3, 4, or 5) using both the studied method and criterion method, and *S* is the total number of raw score points on the New Form.

Results

Sample and Test Characteristics

The data used in the current study were originally collected under the CINEG design and transformed to approximate the RG design using two sampling methods, which resulted in the Matched Samples and Single Population datasets. Both sets of artificial data were treated as though they were collected under the RG design, and the standardized mean difference (i.e., effect size) was computed using common item scores (Table 4). Even though the effect sizes were less than the target effect size of ± 0.05, it was still possible that the distributional shapes of the pseudo-groups differed. Therefore, histograms were used to visually inspect the similarities between the pseudo-groups (Figure 1). The distributional shapes of the pseudo-groups appeared similar except for the Matched Sample Spanish dataset.

Form characteristics. The AP Spanish Language and English exams were made up of MC and FR items. Form characteristics can be found in Tables 5 and 6 for the Matched Samples and Single Population datasets, respectively.

Based on the frequency histograms (Figure 1) and descriptive statistics (Tables 5 & 6) it was apparent that there were far fewer examinees at the upper and lower ends of the composite score scale in both subjects. In order to avoid interpreting equating results that were based on few examinees, a general rule was applied such that results were constricted to score ranges that had a minimum of 10 examinees. This resulted in truncated score ranges in comparison to those listed in Tables 5 and 6. For the Single Population datasets, the truncated raw score scales for Spanish and English were 35 – 131 and 36 – 120, respectively. For the Matched Samples datasets, the raw score scales after truncation were 45 – 131 and 27 – 115 for Spanish and English, respectively.

Test blueprints. Test specification tables were used to inform confirmatory-based dimensionality assessment procedures and to assign items to group-specific factors under the Bifactor MIRT model. Items were classified according to format and content subdomain, each considered separately. The test specification tables for Spanish and English can be found in Tables 7 and 8, respectively. For Spanish, item format and content subdomain were highly related as the listening and reading tasks were in MC format and writing and speaking tasks were in FR format.

Dimensionality Assessments

The Matched Samples data were used to assess form dimensionality in both subjects. First, principal components analysis (PCA) was performed to test the UIRT assumption of unidimensionality. Scree plots containing the eigenvalues can be seen in Figures 2 and 3 for Spanish and English, respectively. In general, the scree plots for both subjects suggested that a two dimensional solution may provide a better fit than a one dimensional solution. According to Reckase, (1979), the unidimensionality assumption is supported when the first eigenvalue accounts for 20% or more of the total variance, which was not found for either subject.

Poly-DIMTEST was also used to test the assumption of essential unidimensionality. In poly-DIMTEST, three different approaches were used. First, a confirmatory approach using item format was run, followed by a confirmatory approach using content subdomain, and lastly, an exploratory approach was run. The probabilities (i.e., *p*-value) associated with the Poly-DIMTEST test statistic can be found in Table 9. The Poly-DIMTEST test statistics indicated that essential unidimensionality held for the most part across both exams. Overall, the *p*-values associated with the Poly-DIMTEST test statistic widely varied depending on which subset of items defined AT1, and suggest results were likely unstable.

Disattenuated correlations were also used to investigate the dimensional structure of the test forms and values can be found in Table10. Results indicated that the MC and FR items for Spanish and English were more likely measuring somewhat different constructs as the disattenuated correlations were in the mid 0.80's and low 0.90's. For English, it appeared as though item format contributed to the multidimensional structure of the exam slightly more than content subdomain, as the correlations were slightly lower. Moderately sized (i.e., mid 0.80's) disattenuated correlations were associated with content for Spanish and tended to be smaller for the Old Form. For Spanish, the smallest disattenuated correlation ($\rho = 0.66$) was found between Reading (C2) and Speaking (C4).

Model Fit

Using the Matched Samples datasets, a series of categorical exploratory and confirmatory factor analyses were conducted to evaluate the structure of the data and the appropriateness of the (M)IRT models used in this study. Across subject exams and for both New and Old forms, a series of 1-, 2-, 3-, and 4-factor EFA models were fitted and AIC and BIC fit indices were computed for each. The AIC and BIC indices for these analyses can be found in Tables 11 and

12, respectively. For the AIC and BIC, smaller values indicate better fit, however no absolute criteria exist in the literature. For both forms of Spanish and English, the AIC behaved as expected such that it favored the EFA higher dimensional solutions. However, the BIC favored the two dimensional EFA solutions for English and the four dimensional solution for Spanish.

The AIC and BIC favored the format-based CFA model for Spanish, but favored the content-based CFA for English (Tables 11 and 12). Considering the EFA and CFA models together, and based on the AIC and BIC values, the highest dimensional EFA solution was favored overall for both forms of Spanish and English.

Fitted Distributions

The fitted distributions were virtually the same for the Old and New form samples, and as a result, only graphical results for the New form samples are provided.

Traditional equipercentile method. For the Matched Samples datasets, traditional equipercentile equating with log-linear presmoothing served as the criterion, and was included as one of the studied equating methods for the Single Population datasets. Presmoothing values were selected based on a series of chi-square difference tests and inspection of raw and fitted distributions. The presmoothing values used in the current study can be found in Table 13.

The empirical and smoothed distributions for the New and Old Form pseudo-groups can be found in Figures 4 to 7. Also in those figures are the fitted form distributions for the UIRT and MIRT models, which are discussed in greater detail next. Overall, the presmoothed distributions fit the empirical distributions very closely for the Old and New forms, across both subjects. For both subjects, the log-linear presmoothed distribution fit the empirical distributions better or equally as well as some of the (M)IRT models, but also appeared less normal.

For the Spanish Matched Samples and Single Population datasets, the traditional method with presmoothing fit the observed distributions most closely, but the (M)IRT methods also appeared to fit reasonably well. One exception was seen with the fitted distribution resulting from the full MIRT 4D orthogonal model, which actually did not appear to fit well for the New or Old form samples.

For the English Matched Samples and Single Population datasets, the fitted distributions for the traditional presmoothed, UIRT, BF, and full MIRT methods all provided a close approximation to the empirical distribution.

Equating Relationships for Single Population Datasets

For the Single Population datasets, there were no difficulty differences across forms and so, theoretically, a score on the pseudo New Form should be equivalent to that same score on the pseudo Old Form. This situation has also been referred to as "equating a test to itself" (Kolen & Brennan, 2014). Equating results were evaluated using graphical displays of raw and scale score equating relationships, DTM plots, along with weighted RMSDs between the (M)IRT method and the criterion Identity method.

Spanish Language. Figure 8 depicts the raw score equating relationships for Spanish using all studied methods. The two reference lines at ± 0.5 indicate the DTM standard for Old Form equivalents. The equating relationship for the full MIRT orthogonal 4D method was most similar to the Identity criterion, and fell within the DTM boundaries for raw and scale scores across the entire score scale. The method most different from the Identity criterion was the traditional equipercentile and UIRT methods. The equating relationship patterns for the remaining (M)IRT methods were very similar to one another.

The scale score equating relationships can be found in Figure 9 for Spanish for all the studied equating methods. The general patterns of the equating relationships were very similar across all methods. For scale scores, the full MIRT orthogonal 4D method came closest to the Identity relationship and was within the DTM boundaries across the entire score scale. The scale score equating relationships for the remaining methods were within the DTM boundaries for raw composite scores of approximately 70 or greater.

English Language. Figure 10 depicts the raw score equating relationships for English using all studied equating methods. The equating relationship for the (M)IRT methods were most similar to the Identity criterion, but fell outside the DTM boundaries for the lower half of the score scale. The traditional equipercentile and UIRT methods were most different from the Identity criterion. Overall, the equating relationship patterns for the (M)IRT methods were very similar to one another.

The scale score equating relationships for English can be found in Figure 11 for all studied methods. The general patterns of the equating relationships were very similar across all methods, with the exception of the traditional equipercentile method, which diverged near the upper end of the score scale. The scale score equating relationship for all (M)IRT methods were within the DTM boundaries, with the exception of the full MIRT 2D methods, which fell outside

the DTM boundaries for raw composite scores greater than 95. The equating relationship for the traditional equipercentile method fell outside the DTM boundaries for raw composite score between 25 and 40 and between 112 and 115.

Equating Relationships for Matched Sample Datasets

The equating criterion used with the Matched Sample datasets was the traditional equipercentile method with log-linear presmoothing. Equating results were evaluated with regard to graphical displays of raw and scale score equating relationships, difference plots and weighted RMSDs between the (M)IRT method and the criterion method.

Spanish Language. The estimated raw score equating relationships for Spanish can be found in Figure 12 for all studied methods. The general pattern for the estimated equating relationships was similar across for all studied methods. However, the relationships for the full MIRT Orthogonal 4D, Orthogonal 2D, and Correlated 4D methods were noticeably more similar to one another. Furthermore, the relationships for the full MIRT Correlated 2D, BF-content, BF-format, UIRT, and traditional equipercentile methods were most similar to one another.

Differences between the Old Form raw score equivalents from the traditional equipercentile method and the (M)IRT methods can be found in Figure 13. From Figure 13, it can be seen that the differences between the equipercentile and the UIRT, BF-Content, BF-Format, and full MIRT Correlated 2D methods are generally less than the DTM standard of ± 0.5 points. Actually, the UIRT method was the only equating procedure that remained inside the DTM boundaries across the entire scale. However, the differences between the equipercentile method and the full MIRT 4D methods and full MIRT 2D orthogonal method are greater than the DTM standard across the entire score scale.

Differences between the Old Form scale score equivalents from the traditional equipercentile method and the (M)IRT methods can be found in Figure 14. The same trend that was found for raw scores was also found with scale scores, with a few exceptions. First, the full MIRT correlated 2D method was no longer entirely within the DTM boundaries, particularly at the lower and upper score ranges. Second, the general pattern for the full MIRT 2D correlated method mirrored those of the other full MIRT methods. This was actually found for the raw score differences, but was less pronounced, and thus harder to identify. Third, the full MIRT 4D methods and orthogonal 2D method were within the DTM boundaries for New Form scale scores near the middle of the scale.

English Language. The estimated raw score equating relationships for English can be found in Figure 15 for the various (M)IRT methods studied. The general shapes of the equating relationships were similar across all methods. It is important to note that four dimensional MIRT methods were not included for English it had two content subdomains and two item formats.

Raw score differences between the traditional equipercentile method and the (M)IRT methods for English, can be found in Figures 16 for all methods. In general, the BF-Format performed most differently from the traditional equipercentile method and had differences greater than the DTM standard for Old Form raw scores between 55 and 115. Differences between the traditional equipercentile method and all MIRT methods were greater than the DTM standard only near the low end of the scale, and were generally within the DTM boundaries for the majority of the score scale.

Differences between the scale score equivalents from the traditional equipercentile method and the (M)IRT methods can be found in Figure 17. The same general trend was observed for scale score equivalents as was found for the raw score equivalents.

Weighted Root Mean Square Differences for Old Form Equivalents

In order to quantify differences between the equating criterion and the studied methods, a weighted Root Mean Square Difference (wRMSD) statistic for Old Form raw and scale score equivalents was computed. RMSDs were weighted using the sampled New Form frequencies.

Single Population datasets. For the Single Population datasets, the wRMSD for raw and scale scores for Spanish and English can be found in Table 14. In general, there was a positive relationship between difference plots and wRMSDs, for both subjects. The full MIRT 4D orthogonal method led to the smallest wRMSD for Spanish raw and scale score equivalents, whereas the biggest difference was found with the traditional equipercentile method, followed closely by the UIRT and both BF methods. The smallest wRMSDs for English raw and scale scores were generally found with the Bifactor and full MIRT 2D methods. For English, the largest wRMSDs for raw and scale scores were found with the UIRT method.

Matched Sample datasets. For the Matched Samples datasets, the wRMSDs for raw and scale scores for Spanish and English can be found in Table 15. For Spanish raw score equivalents, the smallest wRMSD was found for the UIRT, followed by the BF-Content, and full MIRT 2D correlated method, while the largest corresponded to the full MIRT 4D orthogonal method. However, for the Spanish scale score equating relationships, the smallest wRMSDs

corresponded to the UIRT, BF-Format, and BF-Content methods, and the largest still corresponded to the full MIRT 4D orthogonal method. For English raw and scale score equivalents, the smallest wRMSDs were found for the UIRT and BF-Content methods and the largest was found for the BF-Format method.

AP Grade Agreements

Since the data used in this study were originally collected with the intent to classify students into one of five grade levels, it was important to look at classification agreements among all studied methods and their respective criterions. For each studied equating method, the overall exact agreement percentage (EAP) was computed as the percentage of New Form raw scores that equated to the same Old Form AP Grade equivalent as the respective criterion. The percentage was unweighted and aggregated across the entire score scale.

Single Population datasets. For the Single Population datasets, AP Grade agreement percentages were generally high and ranged between 97.8% and 100% and between 98.8% and 99.1% for Spanish and English, respectively (Table 16). For both subjects, the traditional equipercentile method resulted in the lowest (albeit still relatively high) AP Grade agreements. For Spanish, perfect agreement was obtained for the full MIRT 4D orthogonal method. For English, the UIRT and MIRT methods resulted in the same agreement percentages (i.e., 99.1%), whereas the agreement percentage for the traditional equipercentile method was slightly lower.

Matched Samples datasets. For the Matched Sample datasets, AP Grade agreements were more variable than the Single Population datasets and ranged between 94.8% and 100% and between 97.5% and 100% for Spanish and English, respectively (Table 16). For Spanish, perfect agreement was obtained for BF-Content method. For English, the UIRT and BF-Content methods resulted in perfect agreements with the criterion. Overall, agreement percentages were more consistent for English than Spanish.

Discussion

The purpose of this study was to build upon the existing MIRT equating literature by introducing a full MIRT observed score equating method for mixed-format tests and compare its performance with the traditional equipercentile, UIRT observed score, and BF observed score equating procedures. Specifically, the research objectives were to a) present observed score equating methods for mixed-format tests using a full MIRT framework, b) compare the differences between the full MIRT, BF, UIRT, and traditional equipercentile methods, c)

compare differences in the full MIRT equating results when correlations between latent traits were specified as mutually orthogonal versus when they were freely estimated, and d) examine whether dimensionality associated with item format versus content subdomain influenced equating relationships for the BF method.

For any comparative study, there is a need for a criterion or baseline to which studied methods can be compared. The importance of this criterion cannot be overstated, and affects all subsequent interpretations, which is why it is discussed first. Next, a discussion concerning the dimensionality of the AP Exams is presented, and is followed by a detailed comparison of the studied equating methods. A discussion of the limitations and future directions are presented last.

Importance of the Equating Criterion

Since all analyses involved real data in this study, the true equating relationships were unknown, and therefore, comparisons between the studied equating methods were made with respect to an equating criterion. For the Single Population and Matched Sample datasets, the Identity and presmoothed traditional equipercentile method served as criteria, respectively.

For both AP subjects, the conclusions made were completely different when the criterion was the Identity line versus when it was the traditional equipercentile relationship. The comparisons made with the Identity criterion agreed with the overall findings from the dimensionality assessments, which supported its use as a criterion. Whereas, comparisons made with the traditional equipercentile criterion did not agree with the dimensionality assessment results nor with the equating results obtained by using the Identity criterion. This finding was not surprising since the traditional equipercentile equating relationships were typically most different from the Identity criterion in the Single Population datasets. Since comparisons differed so much depending on the criterion used, it was difficult to make general comparisons using both criteria simultaneously. For this reason, and because of the congruence between the dimensionality assessments and the results using the Identity criterion, more emphasis is placed on results found with the Single Population datasets.

It is important to point out that, for the Single Population datasets, even though both groups were sampled from the same population, there was evidence of slight group differences due to the use of samples. Therefore, the Identity line did not accurately portray the differences that were to be equated. However, this does not imply a problem with the Identity criterion, but instead is more of a sampling issue. As a result, small departures from the Identity equating line

may not be indicative of random or systematic error but, instead, may represent actual group differences. Therefore, it could not be said with certainty that departures from the Identity line reflect error.

Even though the traditional equipercentile method was not a perfect substitute for the true equating relationship, it was the best available option for the Matched Samples datasets and was useful for making comparative evaluations. Therefore, making statements with regard to the accuracy of one equating method over another was generally avoided in this study.

Comparison of Equating Methods: Research Objective #2

A basic and all-encompassing objective of this study was to compare the full MIRT observed score equating methods to the traditional equipercentile, UIRT, and BF methods. In order to avoid redundancy, comparisons in this section are made with respect to only the raw score estimated equating relationships, unless stated otherwise.

Single Population Datasets. For the Single Population datasets, slight deviations from the Identity line were expected due to random sampling error. For the most part, the estimated equating relationships appeared to take on the same shape as the Identity relationship, but some were bumpier than expected.

With Spanish, the estimated equating relationships were similar to one another, with the exception of the full MIRT 4D orthogonal and traditional equipercentile methods. The full MIRT 4D orthogonal relationship was most similar to the Identity relationship, and was the only method to remain within the DTM boundaries across the entire score scale. However, the pattern of item discriminations on did not agree with the test specifications. The traditional equipercentile relationship was the most curvilinear and among the most different from the Identity relationship, which was unexpected. The differences seen with the traditional equipercentile method are less likely to be due to group differences since the shape of the ability distributions were very similar for the Old and New Form groups. Also, the UIRT relationship was considered similar to the Bifactor and full MIRT (excluding the 4D orthogonal method) relationships for the majority of the score scale, but not at lower score ranges. Specifically, the UIRT relationship was lower than the Identity relationship by approximately two raw score points on the Old Form scale for lower score ranges. This last finding may indicate that examinees with lower abilities are more variable with respect to the latent dimensions being measured by the test items, in comparison to higher ability examinees.

For English, the equating relationships for the Bifactor and full MIRT methods were generally very similar to one another, while the relationships for the UIRT and traditional equipercentile showed somewhat unique patterns. The UIRT relationship followed the same pattern as the MIRT methods but was slightly more curvilinear. The relationship for the traditional equipercentile method was the most curvilinear of all the methods and was most different from the Identity criterion. While none of the studied methods produced equating relationships that were consistently within the DTM boundaries, all differences were within approximately ± 1.0 raw score points.

Matched Samples Datasets. For the Matched Samples datasets, the shapes of the estimated equating relationships were used to compare the studied equating methods. In general, different conclusions were made for Spanish and English.

For Spanish, the estimated equating relationships for all studied methods had the same general shape; however there were two clear groupings. In the first group were the estimated equating relationships for both four-dimensional full MIRT models and the two-dimensional orthogonal full MIRT model. In the second group were the traditional equipercentile, UIRT, both BF methods and the two-dimensional correlated full MIRT method. The equating relationships in the second group suggested smaller differences between forms at low abilities and larger differences between forms at high abilities, in comparison to the methods in the first group. At this time, it is difficult to determine what caused the two general groupings; however, it is apparent that the group including the UIRT and Bifactor methods represents less complex models.

Unlike with the Spanish exams, the equating relationships for English were all very similar, and no clear groupings emerged. All equating methods produced the same general pattern, and would result in Old Form equivalents within one raw score point of one another, if used operationally.

Summary. In general, it seems that the shapes of the equating relationships for the traditional equipercentile, UIRT, Bifactor, and full MIRT methods had very similar features with a few exceptions. For Spanish and English Single Population datasets, the traditional equipercentile method resulted in an estimated equating relationship that was generally most different from the Identity criterion. The relationships found by using the traditional equipercentile method were also generally more curvilinear than the other methods. There have

been mixed findings in the literature on whether traditional or IRT equating methods are more robust to group differences (Eignor, Stocking, & Cook, 1990; Schmitt et al., 1990). Therefore, it is difficult to hypothesize the extent to which the presence of group differences affected equating results. Even though examinees were drawn from the same population (i.e., the original Old Form), the act of sampling 3,000 to create the pseudo-forms may have resulted in group differences that were large enough to impact the traditional equipercentile method but not the (M)IRT methods. On the other hand, the traditional equipercentile equating relationship may represent the small group differences more accurately than the (M)IRT methods. At this point, it is difficult to make that judgment since the population equating relationship is unknown.

Effect of Latent Trait Structure on Full MIRT Equating: Research Objective #3

In the MIRT literature it is typically more common to find latent traits treated as orthogonal because doing so simplifies procedures such as scale linking. However, this may not always be realistic, especially with educational data. Therefore, this study investigated the effect of treating latent traits as orthogonal versus correlated for the full MIRT methods. In general, the differences in equating relationships that were associated with the structure of the latent abilities were not consistent across the Single Population and Matched Samples datasets.

Single Population Datasets. For Spanish Language, the full MIRT orthogonal methods led to equating relationships that were generally more similar to the criterion relationship for raw and scale scores. For the four-dimensional models, the equating relationship patterns for the orthogonal and freely correlated conditions were clearly different from one another.

For English, specifying the latent traits as orthogonal or freely correlated for the two-dimensional full MIRT methods had little impact on the raw and scale score equating relationships. Both approaches led to equating relationships that were very similar to the Identity criterion.

Matched Samples Datasets. For Spanish Language and English, the equating relationship patterns were very similar regardless of whether the latent traits were allowed to correlate or not. For Spanish, the raw and scale score equating relationships from the orthogonal methods were slightly lower than those when trait correlations were freely estimated.

Summary. In general, the treatment of latent traits as orthogonal versus correlated did not appear to have a large influence on the raw and scale score the equating relationships. One exception was seen with the Spanish Single Population datasets, but this difference was with

respect to the four-dimensional models only. For the English Single Population and Matched Samples datasets, the full MIRT correlated and orthogonal equating relationships intersected one another too frequently to be considered different.

Modeling Dimensionality by Content Versus Item Format: Research Objective #4

In this study, when the Bifactor model was used, group-group specific dimensions corresponded to either item format (labeled BF-Format) or content subdomain (labeled BF-Content) and differences were evaluated. Lee and Lee (2014) presented a Bifactor observed score equating method where item format represented the group-specific factor; however, there are no equating studies that defined group-specific factors according to content. Since the group-specific dimension in the Bifactor model represents the residual variance not accounted for by the general dimension, large differences between approaches were not expected.

Single Population Datasets. For Spanish and English, the BF-Content and BF-Format equating relationships patterns for raw and scale scores were very similar, overall. Slight differences between the BF-Content and BF-format equating relationships were seen for Spanish, but only at the upper end of the score scale.

Matched Samples Datasets. In general, more discrepancies were seen between the BF-Content and BF-Format estimated equating relationships for the Matched Samples datasets. This trend may be related to the fact that there were larger group differences in these datasets. For Spanish, the Bifactor equating relationships were similar in general, but crossed one another near the middle of the score scale. Furthermore, the BF-Content relationships appeared slightly bumpy in comparison to the BF-Format relationships, which is likely due to greater number of group-specific traits.

The biggest difference between the BF-Content and BF-Format estimated equating relationships was found for the English exams. Specifically, the BF-Content relationship was consistently parallel to, and below the BF-Format relationship, across the entire score scale. This finding may suggest that the forms and/or examinee groups differed more with respect to item format effects at lower end of the scale and differed more on content subdomain at the upper end.

Summary. The effect of defining group-specific factors according to item format in comparison to content subdomain varied depending on which datasets (i.e., Single Population or Matched Samples) were being referenced. However, across both datasets, the BF-Content equating relationships were sometimes slightly bumpy whereas this was never found for the BF-

Format relationships. This finding may be related to the increase in the number of parameters that needed to be estimated in the former method. It should also be noted that in order to estimate the marginal total score distribution under the BF-Content method, a reduced number of quadrature points (i.e., 11 as opposed to 25) was necessary in order to avoid convergence problems. Therefore, the bumpy pattern is likely related to the need to use fewer quadrature points.

Even with the few instances where the BF-Content and BF-Format equating relationships appeared different, the majority of the time differences were trivial. This is likely related to the fact that under the Bifactor model, the majority of the variance is attributed to the general factor, which remained constant. If there is not a lot of residual variance left over, then modeling it according to content versus format may not make that much of a difference. Therefore, if observed score Bifactor equating is desirable, then it may be better to specify the group-specific factors according to item format, which typically does not have many levels.

Limitations and Future Considerations

Like all studies, this study is not without several limitations that are worth mentioning. Several of the limitations of this study can be linked to the fact that analyses were conducted using real data. While this is not a limitation in and of itself, it did invoke specific challenges.

The biggest challenge associated with the exclusive use of real data was that the true equating relationships for the AP Exams were not known. This resulted in the need to rely on criteria that contained some error. With this in mind, an attempt was made to quantify the amount of error in one of the equating criteria – the traditional equipercentile method – by including it in conditions where equating was theoretically unnecessary (i.e., equating a test to itself). As it turned out, the traditional equipercentile method performed quite differently from the Identity criteria and the other methods. However, it is difficult to say which method was actually closest to the population equating relationship since small group differences were present. Currently, the traditional equipercentile method is likely the best available criterion for equating studies using Matched Samples from real data; however further research is needed in this area. Future studies could ameliorate some of the current challenges by including a simulation study and specifying the true equating relationship a priori. One way to do this would be to use a pseudo-form design with large sample sizes and use a single group equating criterion.

The next limitation concerns the design components surrounding the Single Population datasets where the Identity relationship served as the criteria. For example, the pseudo-forms sampled and used with the Identity criterion were limited to a form size of 3,000, which was likely too small and as a result small group differences were present. However, the Identity equating relationship does not take group differences into account and thereby contains random error. Future studies could improve upon this design by using a larger sample size in order to further reduce random noise.

Another obvious limitation in this study was that the AP datasets were not very multidimensional. This posed a serious challenge since a primary objective was to present and evaluate a full MIRT observed score equating method. The AP exams chosen in this study were selected because they showed evidence of being multidimensional, relative to other AP exams not presented in this study. However, even for the chosen subject areas, an argument could be made that the unidimensionality assumption was still not severely violated. Therefore, the equating relationships found in this study, and the comparative judgments made among them, may have been quite different if more multidimensional data were used.

In addition to overcoming the challenges imposed by using real data, the overall design of this study could be improved in several ways as well. First, the manner in which the full MIRT equating methods were carried out was exploratory, such that items were free to load on all latent dimensions. For the majority of the exams, the pattern of item loadings or item discriminations, did not align to the test specifications and were otherwise uninterpretable. This type of exploratory approach is less likely to be carried out in an operational setting given the large amount of time devoted to test development processes.

Last, all MIRT equating procedures are subject to the limitation associated with the parallelism requirement which must be met in order for equating to be considered appropriate. This prerequisite to equating becomes more difficult to verify as the number of latent abilities increases, such as with going from a UIRT to a MIRT framework. In a MIRT framework, there is more than one latent construct to account for and all must have some degree of parallelism across forms. In addition, the relationships between constructs ought to be similar across forms. Therefore, the requirement of parallelism becomes more difficult to verify as the number of latent abilities increases. However, this may be considered less of an issue for the Random Groups design, since groups are assumed to be randomly equivalent on all constructs.

Conclusions

In this study, a full MIRT observed score equating procedure for mixed-format tests was introduced and compared with traditional equipercentile, UIRT, and Bifactor equating methods. For the Bifactor methods, group-specific factors were modeled according to item format or content subdomain, and either two or four dimensions were specified for the full MIRT methods. AP Spanish Language and English Language and Composition exams were used to illustrate the equating procedures and comparisons were made with respect to two different criteria.

Several important conclusions were stated in the current study and are reviewed here. In general, it was found that the multidimensional methods led to more accurate equating for datasets that evidenced some multidimensionality. Specifically, this was found in the wRMSDs and AP Grade agreement percentages for the Single Population datasets.

The scale on which scores are reported influenced comparative judgments made among the studied methods. For example, with Spanish Language Single Population datasets, the full MIRT four-dimensional orthogonal equating relationship was by far, the most similar to the criterion for raw and scale scores. However, the AP Grade equivalents resulting from that method and the UIRT method were the same for over 99% of the score scale. Therefore, on the AP Grade scale, which is operationally the most important scale for examinees, there typically were not large discrepancies between MIRT and UIRT methods. Since this was found with exams that were not very multidimensional, overestimating data dimensionality may not significantly impact equating results for scores derived similarly to AP Grades.

Another important finding was that the Bifactor observed score equating method may not be very sensitive to the manner in which the group-specific factors are defined. There were generally very little differences between the equating relationships for the BF-Content and BF-Format methods, with the exception of the Matched Sample English exam. Therefore, when the Bifactor method is preferred, it may be beneficial to choose the more parsimonious version of the Bifactor model due to practical constraints associated with estimation and convergence issues.

In general, the AP Exams used in this study did not severely violate the UIRT unidimensionality assumption. It is worth mentioning that Spanish and English evidenced the greatest degree of multidimensionality among the 14 available AP Exams that could have been included, yet were still not very multidimensional. This calls into question the costs and benefits associated with researching and developing new MIRT equating methodology. On the contrary,

if the majority of tests are unidimensional, but a small fraction is multidimensional, the costs associated with improving the measurement procedures for even a small fraction of tests has merit.

In conclusion, similarities in equating relationships were usually seen within the UIRT and Bifactor methods and within the full MIRT methods, and in some instances very few differences were found across all methods (excluding the traditional equipercentile). Of importance is that when the unidimensionality assumption is suspect, using a multidimensional equating method appears to provide a reasonable alternative, and is theoretically more defensible. However, conclusions made in this study were based on a small sample of AP exams and should be replicated using exams with different characteristics.

Any time an IRT observed score equating method is considered with mixed-format exams, special attention needs to be given to the dimensionality of the datasets. Based on the results from this study, standard exploratory analyses such as PCA and EFA may not provide useful information concerning the latent dimensional structure, and may be better suited to help judge whether unidimensionality holds. Furthermore, the information provided by PCA and disattenuated correlations appeared to be more reliable or stable than results from Poly-DIMTEST. Based on the results in this study, it is likely more defensible to specify latent dimensions according to test blueprints. With this said, an SS-MIRT or Bifactor equating method may be the best options for equating mixed-format exams that evidence multidimensionality. Taking this approach guarantees that the latent dimensions will be interpretable, which cannot be said for exploratory full MIRT models, such as those used in this study.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*, 716-723.

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*. Reading, MA: Addison-Wesley.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education*, *12*, 383-407.

Brennan, R. L., & Kolen, M. J. (1987a). A reply to Angoff. *Applied Psychological Measurement*, *11*, 301–306.

Brennan, R. L., & Kolen, M. J. (1987b). Some practical issues in equating. *Applied Psychological Measurement*, *11*, 279–290.

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, *37*, 460-481.

Cai, L. (2012). *flexMIRT* (Version 1.88) [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221-248.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, *32*, 79-96.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (ETS Research Report 85-30). Princeton, NJ: Educational Testing Services.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

de Champlain, A. F. (1996). The effect of multidimensionality of IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*, 181-201.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three advanced placement program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (pp. 79-118), Research Report 03-27. Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*, 249-262.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3*, 37-52.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423-436.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items*. Unpublished research note. Iowa City, IA: ACT, Inc.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph No. 2.2*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, G., & Lee, W. (2014). Bi-factor MIRT observed-score equating for mixed-format tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3).* (*CASMA Monograph No. 2.3*). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Li, Y., Li, S., & Wang, L. (September, 2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (ETS Research Report 10-21). Princeton, NJ: Educational Testing Service.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (April, 1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Li, Y. H., & Stout, W. D. (1995). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 452-461.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E., & Carlson, J. E. (1993). *Full-information factor analysis for polytomous item responses*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reise, S. P., Waller, N. G, & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12,* 287–297.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education*, *3*, 53–71.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37,* 329-346.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item respond theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39-49.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*, 177-197.

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

Zhang, J. (1996). *Some fundamental issues in item response theory with applications* (Unpublished doctoral study). Department of Statistics, University of Illinois at Urbana-Champaign.

Zhang, J. M., & Stout, W. F. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

Zhang, J., & Wang, M. (1998, April). *Relating reported scores to latent traits in a multidimensional test*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago, IL: Scientific Software.

Table 1

Scores and Weights of the MC and FR Items for AP Main and Alternate Forms

	Operational Settings			Studied Settings		
Exam	Section	*N* Items (Categories)	Weights	Section	*N* Items (Categories)	Weights
Spanish	MC	34 (0 – 1)	0.8823	MC	34 (0 – 1)	1
Language		36 (0 – 1)	1.2500		36 (0 – 1)	1
	FR	3 (0 – 5)	3.0000	FR	3 (0 – 5)	3
		1 (0 – 5)	6.0000		1 (0 – 5)	6
English	MC	54 (0 – 1)	1.1250	MC	54 (0 – 1)	1
	FR	3 (0 – 9)	3.0556	FR	3 (0 – 9)	3

Table 2

Studied Conditions for Single Population and Matched Samples Datasets

Model	AP Subject	Ability Distribution	Dimensionality	Condition Number
UIRT	Spanish	UVN (0, 1)	N/A	1
	English	UVN (0, 1)	N/A	2
Bifactor	Spanish	MVN (0, **I**)	Item Format	3
		$\rho = 0$	Content	4
		MVN (0, 1)	Item Format	5
		ρ = estimated	Content	6
	English	MVN (0, **I**)	Item Format	7
		$\rho = 0$	Content	8
		MVN (0, 1)	Item Format	9
		ρ = estimated	Content	10
Full MIRT	Spanish	MVN (0, **I**)	2 Dimensions	11
		$\rho = 0$	4 Dimensions	12
		MVN (0, 1)	2 Dimensions	13
		ρ = estimated	4 Dimensions	14
	English	MVN (0, **I**)	2 Dimensions	15
		$\rho = 0$		
		MVN (0, 1)	2 Dimensions	16
		ρ = estimated		
Equipercentile	Spanish	N/A	N/A	17
	English	N/A	N/A	18

Table 3

Calibration Settings Used During Item Parameter Estimation

General Settings	UIRT	Bifactor	Full MIRT
Estimation Method	BAEM	BAEM	BAEM
Max. # of cycles	3,000	3,000	3,000
Convergence criterion	1e-6	1e-6	1e-6
Max. # of M-step iterations	3,000	3,000	3,000
Convergence criterion	1e-6	1e-6	1e-6
# of quadrature points	41	41	41
Range of quadrature points	[-6, 6]	[-6, 6]	[-6, 6]
Prior Setting			
Prior for *a*-parameter	lognormal ($\mu = 0, \sigma = 0.5$)	lognormal ($\mu = 0, \sigma = 0.5$)	lognormal ($\mu = 0, \sigma = 0.5$)
Prior for *c*-parameter	beta ($\alpha = 4, \beta = 16$)	beta ($\alpha = 4, \beta = 16$)	beta ($\alpha = 4, \beta = 16$)

Note. BAEM stands for the Bock-Aitkin Estimation Method. *flexMIRT* uses (α-1) and (β-1) to describe a beta distribution.

Table 4

Effect Sizes for Group Differences in Common Item Scores

	Spanish	English
Matched Samples	-0.019	0.016
Single Population	-0.012	0.027

Note. Negative values indicate the New Form group was higher achieving.

Table 5

Descriptive Statistics for Forms Associated with the Matched Samples Design

	Spanish		English	
Statistic	New Form	Old Form	New Form	Old Form
Scale	0-145	0-145	0-135	0-135
Mean	89.446	90.191	79.841	77.633
Median	92	92	82	78
SD	23.754	21.679	21.468	18.661
Minimum	0	14	13	14
Maximum	144	143	134	131
Kurtosis	-0.044	0.147	-0.363	-0.268
Skewness	-0.434	-0.461	-0.287	-0.228
Reliability	0.875	0.855	0.878	0.856

Note. Scale represents the total range of composite scores. Reliability is defined using stratified alpha where strata represent item format.

Table 6

Descriptive Statistics for Forms Associated with the Single Population Design

	Spanish		English	
Statistic	New Form	Old Form	New Form	Old Form
Scale	0-145	0-145	0-135	0-135
Mean	91.292	90.914	76.057	76.488
Median	93	93	78	78.5
SD	21.332	21.844	20.249	19.994
Minimum	17	13	0	0
Maximum	142	144	129	129
Kurtosis	0.083	-0.031	0.025	0.083
Skewness	-0.460	-0.441	-0.463	-0.446
Reliability	0.855	0.863	0.871	0.867

Note. Scale represents the total range of composite scores. Reliability is defined using stratified alpha where strata represent item format.

Table 7

Test Blueprint for Spanish Language

	Content Subdomain			
Item Format	Content 1	Content 2	Content 3	Content 4
Multiple Choice	34	36	0	0
Free Response	0	0	2	2

Table 8

Test Blueprint for English Language and Composition

	Content Subdomain	
Item Format	Content 1	Content 2
Multiple Choice	41	13
Free Response	2	1

Table 9

Poly-DIMTEST Test Statistic Probabilities for Matched Sample Forms

	Spanish		English	
Approach	New Form	Old Form	New Form	Old Form
CFA Format	0.091	0.077	0.628	0.239
CFA Content				
F1 as AT1	--	--		
F2 as AT1	--	--		
F3 as AT1	0.500	0.500	--	--
F4 as AT1	0.622	0.701	--	--
EFA	0.001	0.816	0.356	0.037

Note. "--" indicates the number of items in the content subdomain exceeded the limits imposed in Poly-DIMTEST.

Table 10

Disattenuated Correlations for Matched Sample Forms

	Spanish		English	
Source	New Form	Old Form	New Form	Old Form
Format	0.890	0.880	0.784	0.794
Content				
C1 with C2	0.813	0.873	0.826	0.819
C1 with C3	0.881	0.875	--	--
C1 with C4	0.886	0.814	--	--
C2 with C3	0.883	0.868	--	--
C2 with C4	0.660	0.657	--	--
C3 with C4	0.951	0.839	--	--

Table 11

AIC Indices for EFA and CFA Fitted Models

	Spanish		English	
Fitted Model	New Form	Old Form	New Form	Old Form
1 Factor EFA	261956.3	259337.5	194033.8	202297.3
2 Factor EFA	257050.3	255550.3	191900.0	200280.1
3 Factor EFA	255981.1	254385.3	--	--
4 Factor EFA	255534.7	253941.3	--	--
Format CFA	261475.3	258976.9	193328.9	201843.1
Content CFA	300649.8	295579.3	192718.1	201016.4

Table 12

BIC Indices for EFA and CFA Fitted Models

	Spanish		English	
Fitted Model	New Form	Old Form	New Form	Old Form
1 Factor EFA	262941.3	260322.5	194844.7	203108.1
2 Factor EFA	258473.8	256973.8	193047.2	201427.3
3 Factor EFA	257837.1	256241.3	--	--
4 Factor EFA	257817.1	256223.7	--	--
Format CFA	262466.4	259968.0	194145.7	202659.9
Content CFA	302007.3	296936.7	193522.9	201821.2

Table 13

Presmoothing Parameters Used for the Old and New Form Samples

Dataset	Spanish	English
Matched Samples	4, 4	5, 5
Single Population	5, 5	5, 5

Note. The first and second values are smoothing parameters for the New and Old Forms, respectively.

Table 14

Weighted Root Mean Squared Differences Using the Single Population Datasets

	Spanish		English	
Model	Raw Score	Scale Score	Raw Score	Scale Score
UIRT	0.56737	0.23603	0.58791	0.26618
Bifactor Format	0.59559	0.24626	0.49498	0.22287
Bifactor Content	0.58629	0.24366	0.47230	0.21219
Orthogonal 2D	0.50081	0.20721	0.47815	0.21318
Correlated 2D	0.47911	0.24460	0.47467	0.21012
Orthogonal 4D	0.24613	0.11584	--	--
Correlated 4D	0.43743	0.18603	--	--
Equipercentile	0.67512	0.28098	0.52749	0.23445

Table 15

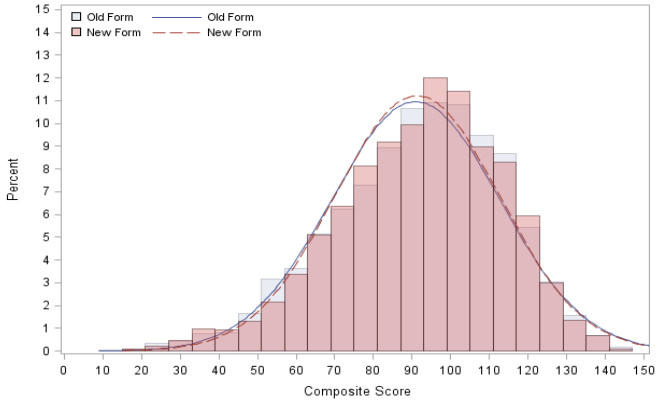
Weighted Root Mean Squared Differences for Raw Scores Using the Matched Samples Datasets

	Spanish		English	
Model	Raw Score	Scale Score	Raw Score	Scale Score
UIRT	0.14301	0.06627	0.14010	0.09080
Bifactor Format	0.32558	0.15972	0.73416	0.39700
Bifactor Content	0.17568	0.10278	0.15680	0.08870
Orthogonal 2D	0.84246	0.76418	0.41444	0.24466
Correlated 2D	0.17637	0.40534	0.36560	0.20586
Orthogonal 4D	0.99176	0.82973	--	--
Correlated 4D	0.73380	0.71567	--	--

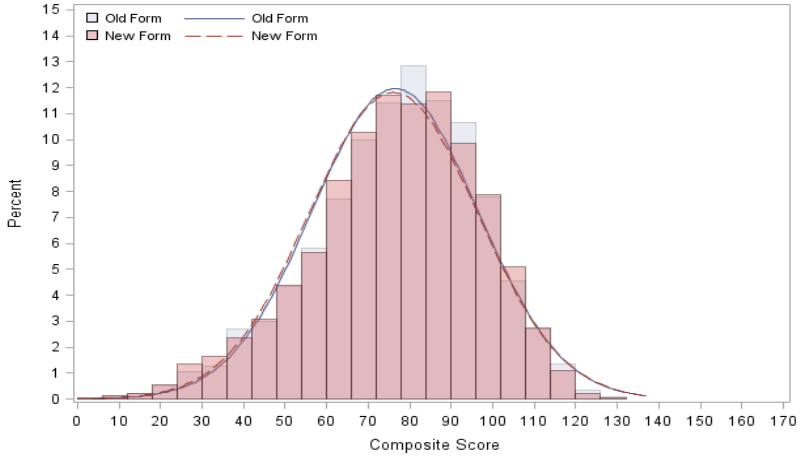
Table 16

Overall Exact Agreement Percentage of AP Grade Classifications for All Datasets

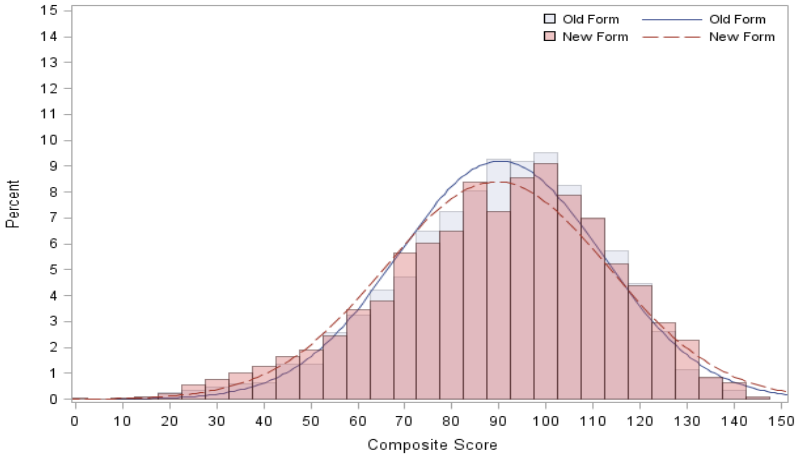
	Single Population		Matched Samples	
Model	Spanish	English	Spanish	English
UIRT	0.99233	0.99066	0.989	1
Bifactor Format	0.97833	0.99066	0.979	0.975
Bifactor Content	0.97833	0.99066	1	1
Orthogonal 2D	0.99233	0.99066	0.965	0.994
Correlated 2D	0.98066	0.99066	0.990	0.992
Orthogonal 4D	1	--	0.948	--
Correlated 4D	0.99233	--	0.965	--
Equipercentile	0.97833	0.98833	--	--



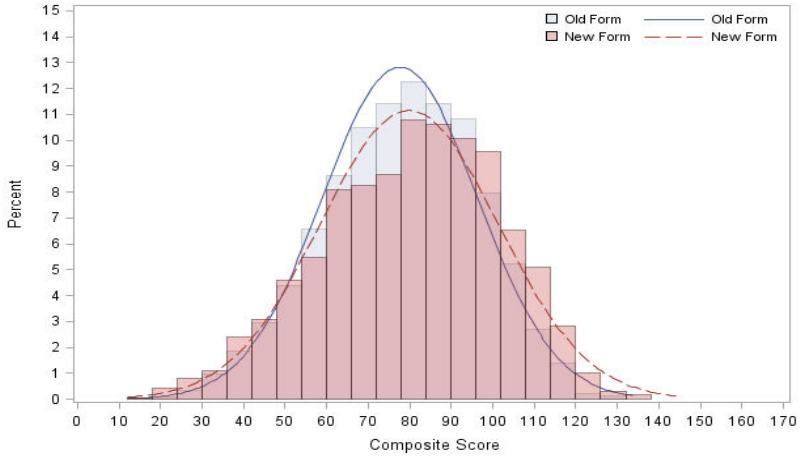
Spanish: Single Population Dataset



English: Single Population Dataset



Spanish: Matched Samples Dataset



English: Matched Samples Dataset

Figure 1. Weighted total score distributions for pseudo old and new form examinees.

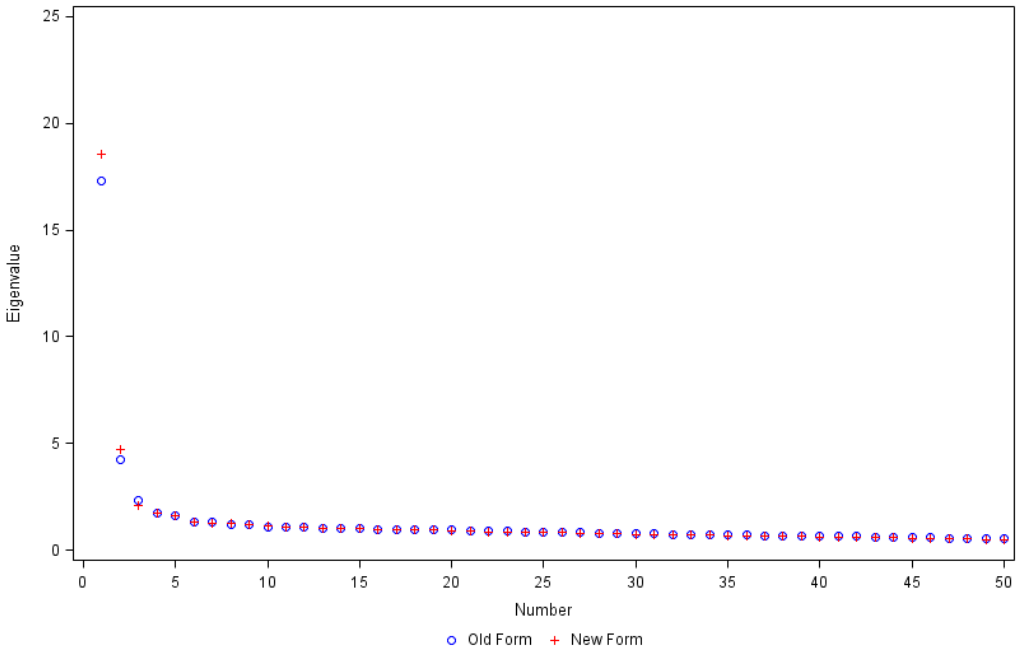


Figure 2. Scree plot of eigenvalues for Spanish using the Matched Samples dataset.

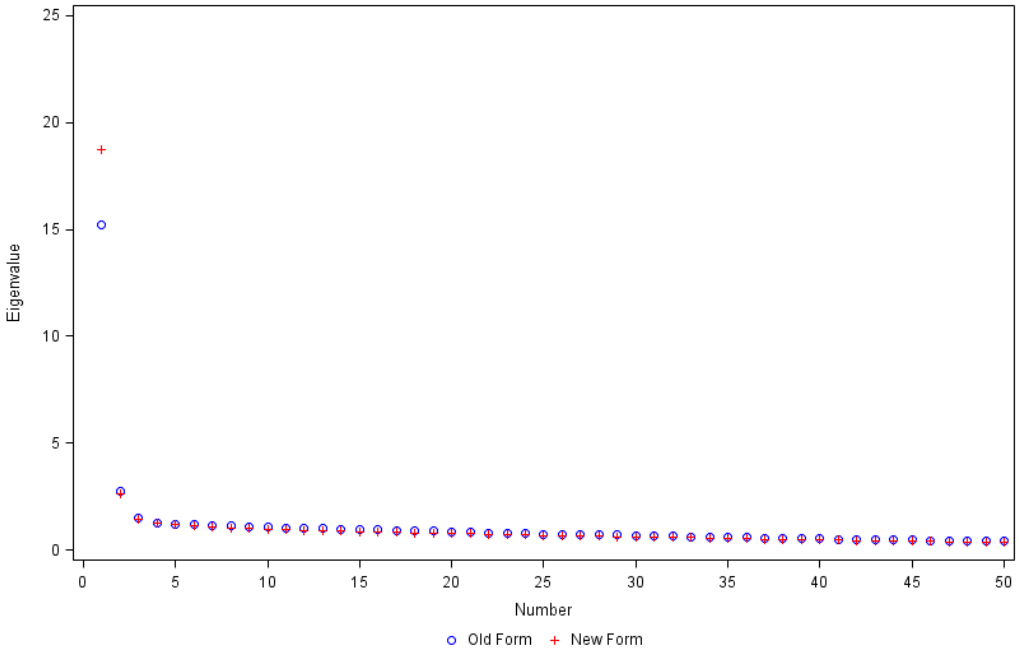


Figure 3. Scree plot of eigenvalues for English using the Matched Samples datasets.

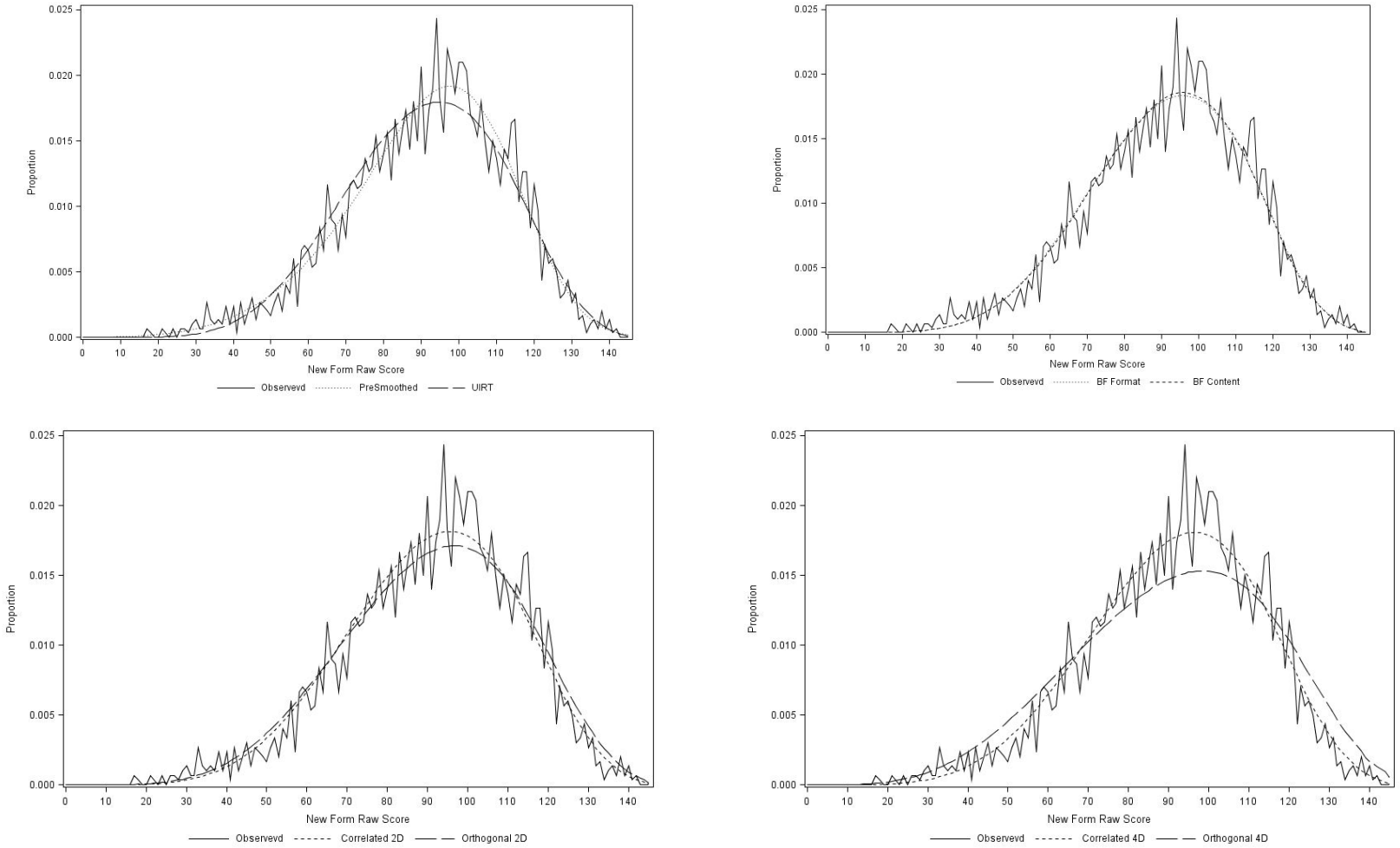


Figure 4. Spanish new form observed and fitted distributions for Matched Samples data.

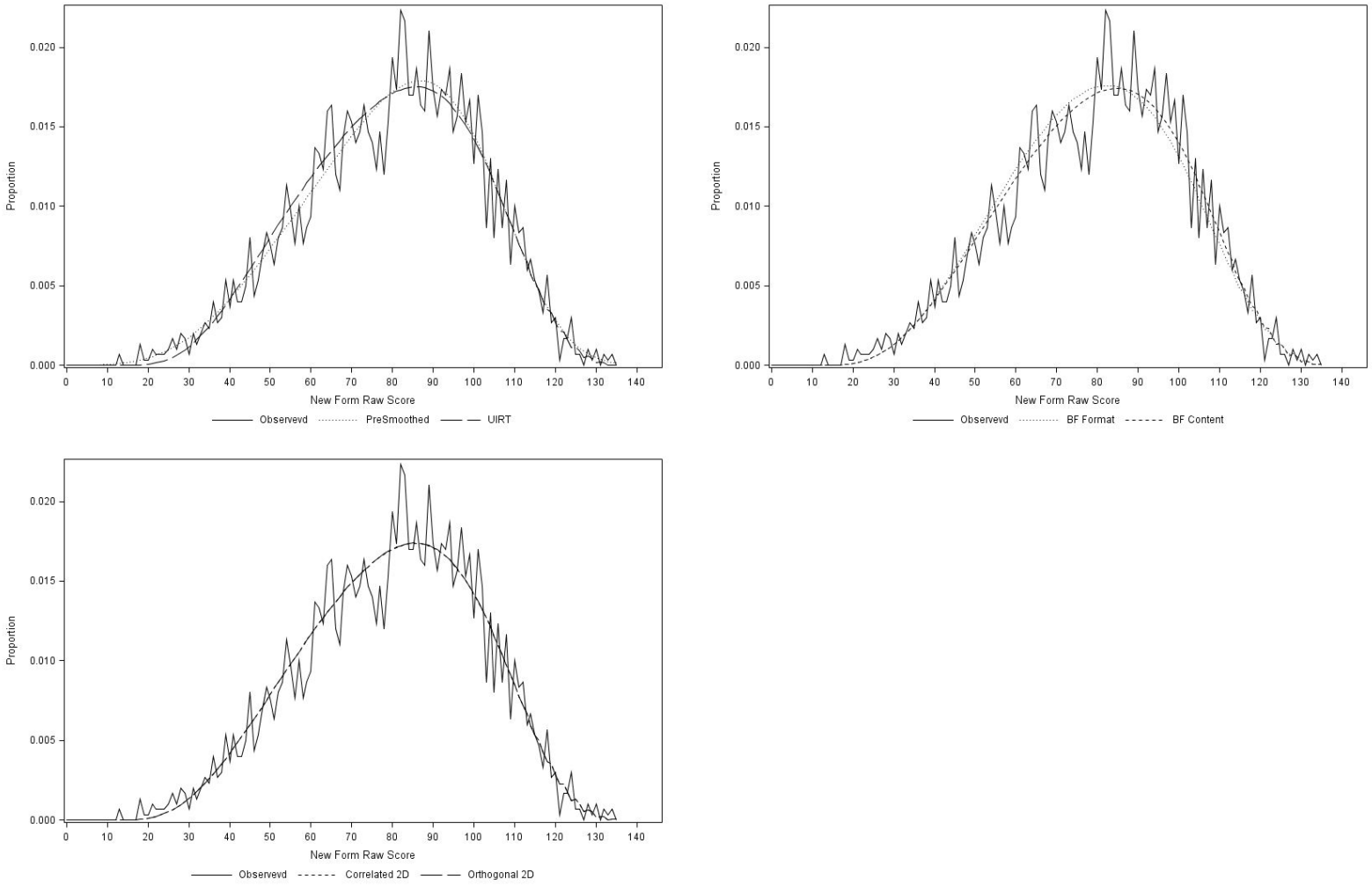


Figure 5. English new form observed and fitted distributions for Matched Samples data.

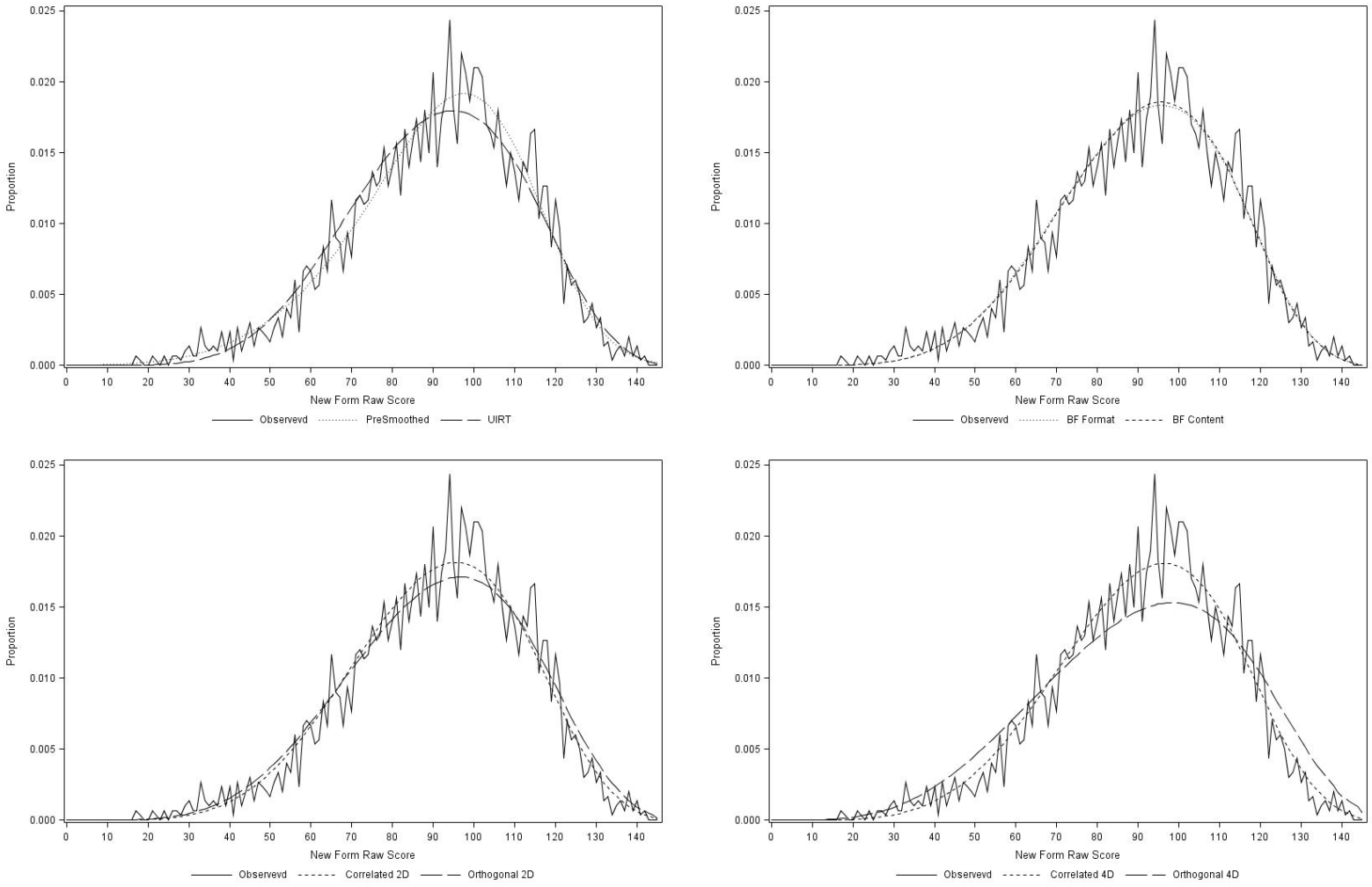


Figure 6. Spanish new form observed and fitted distributions for single population data.

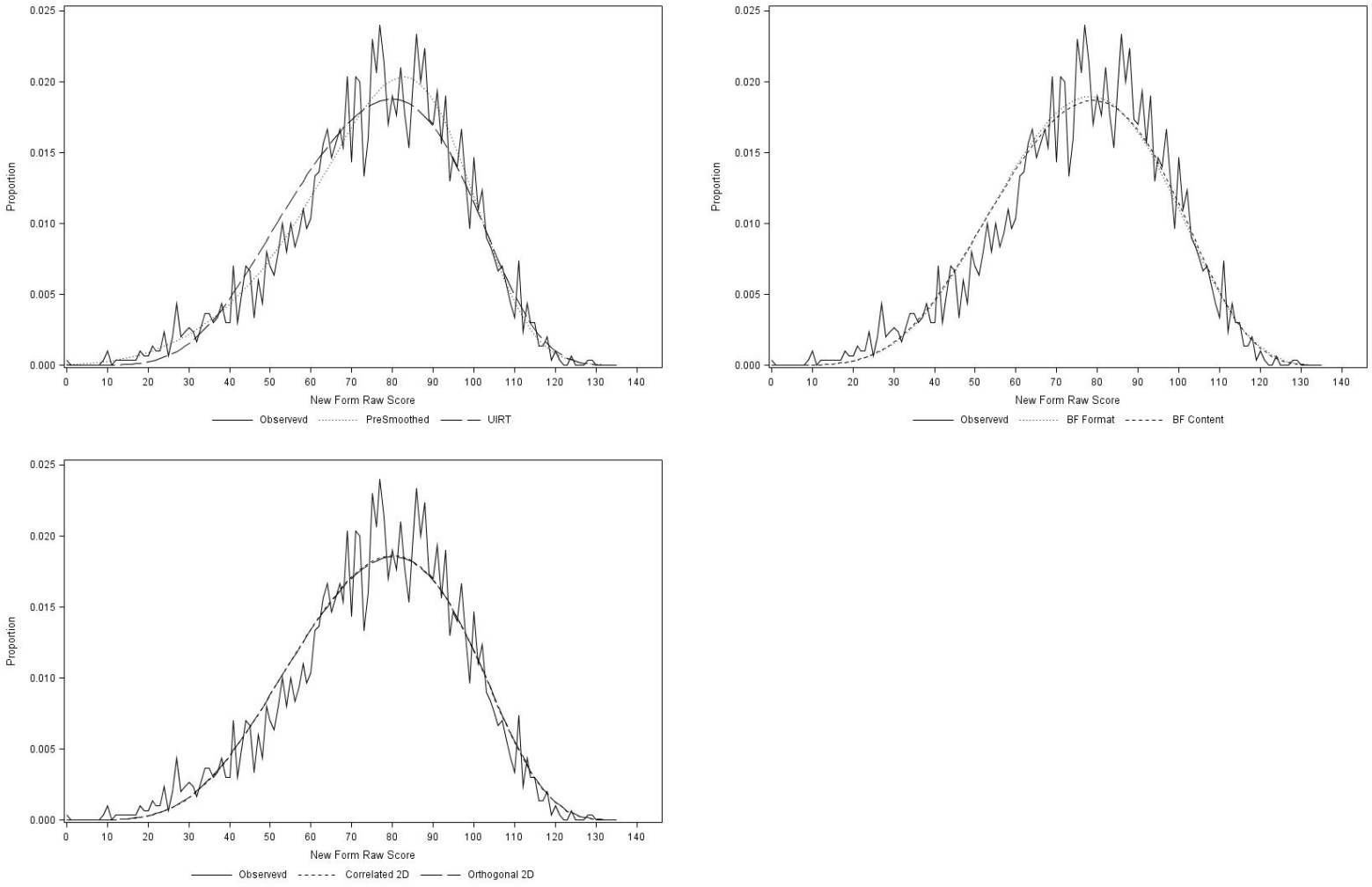


Figure 7. English new form observed and fitted distributions for Single Population data.

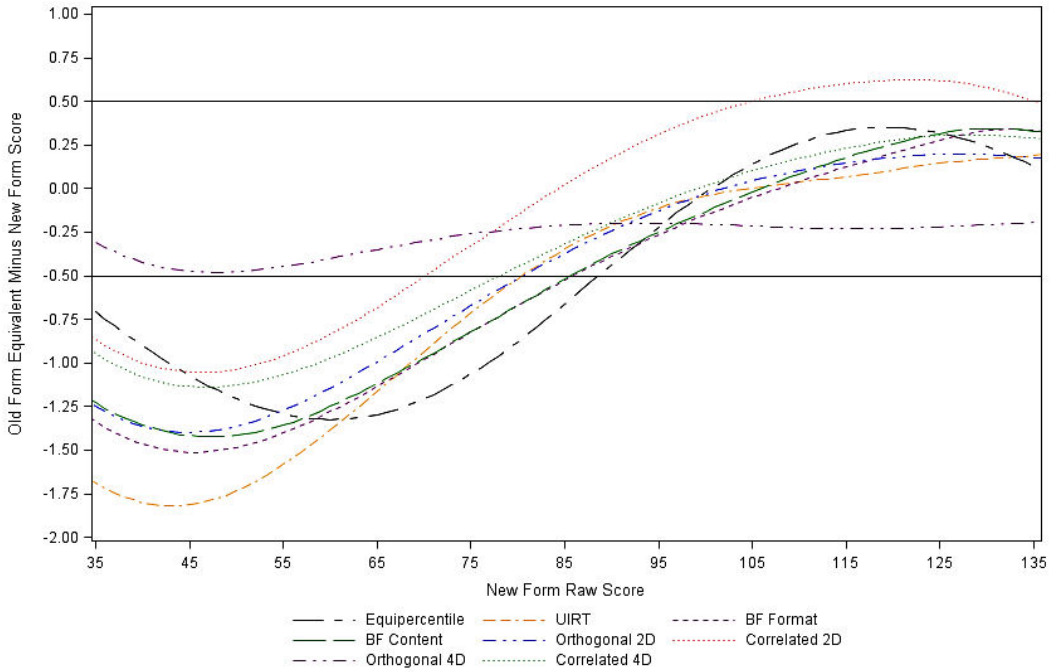


Figure 8. Spanish raw score equating relationships and differences for all methods using the Single Population datasets.

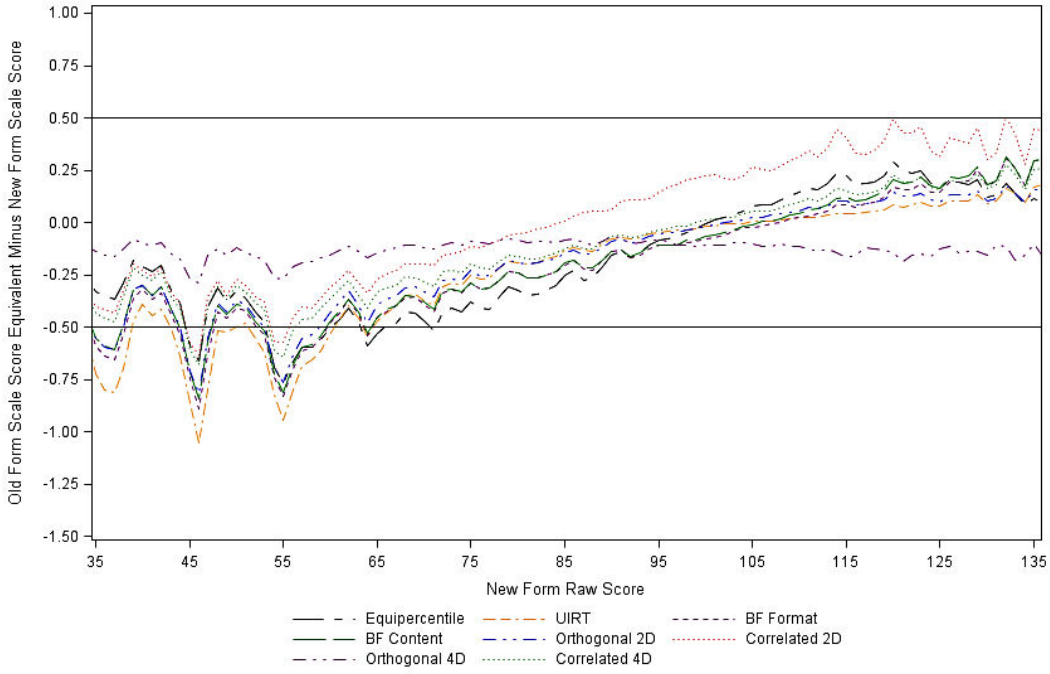


Figure 9. Spanish scale score equating relationships and for all methods using the Single Population datasets.

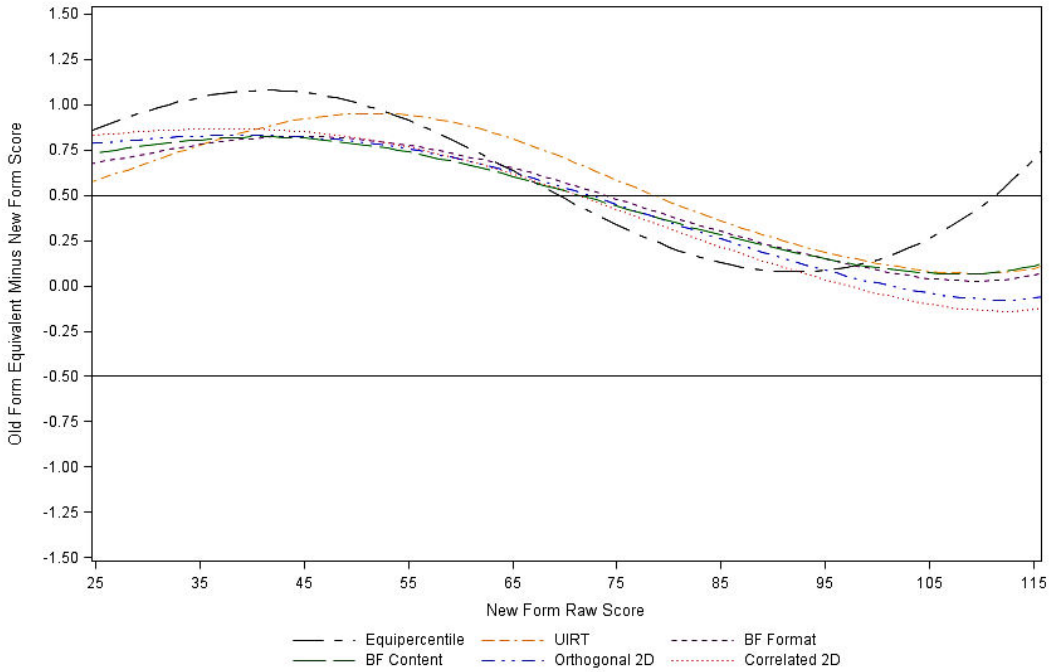


Figure 10. English raw score equating relationships and differences using the Single Population datasets.

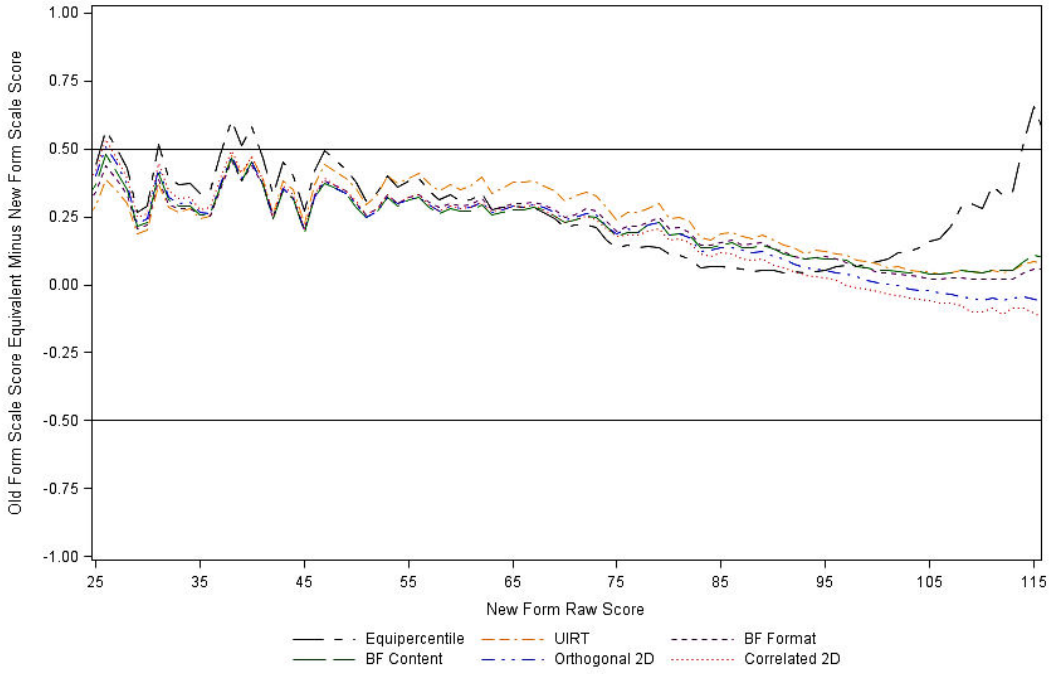


Figure 11. English scale score equating relationships and differences for all methods using the Single Population datasets.

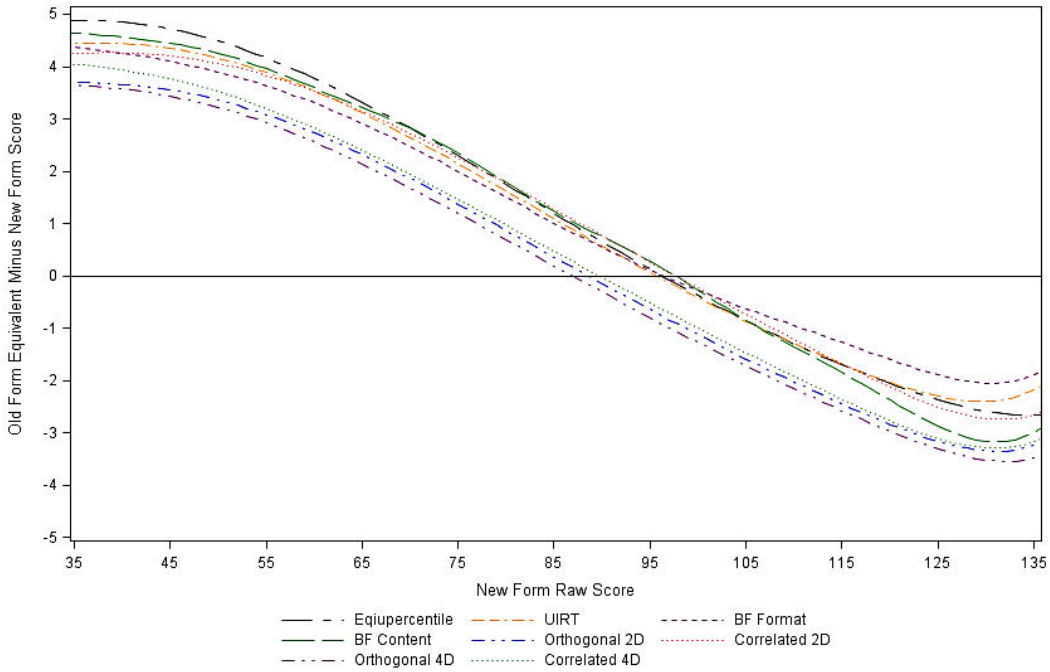


Figure 12. Spanish equating relationships for raw scores using the Matched Samples datasets.

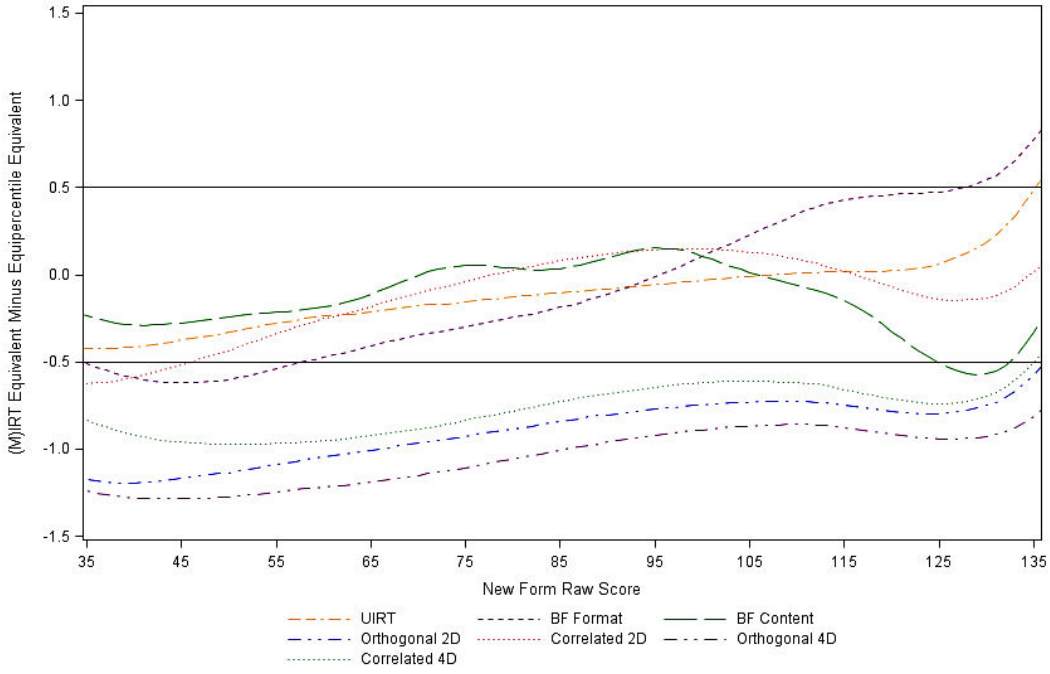


Figure 13. Differences between (M)IRT equivalents and equipercentile equivalents for Spanish raw scores using the Matched Samples datasets.

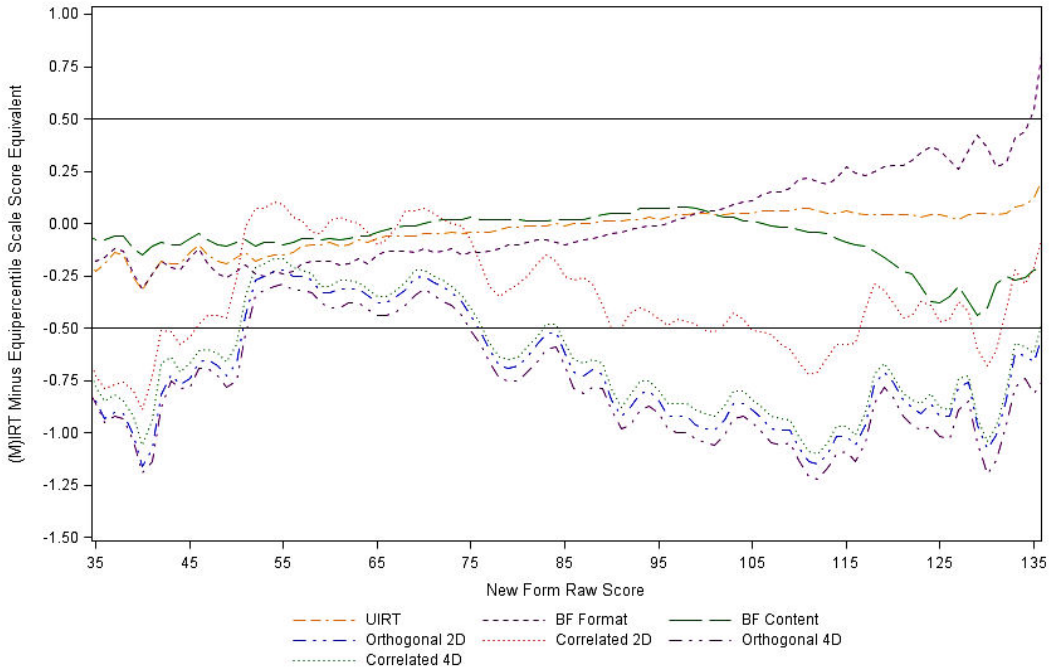


Figure 14. Differences between (M)IRT equivalents and equipercentile equivalents for Spanish scale scores using the Matched Samples datasets.

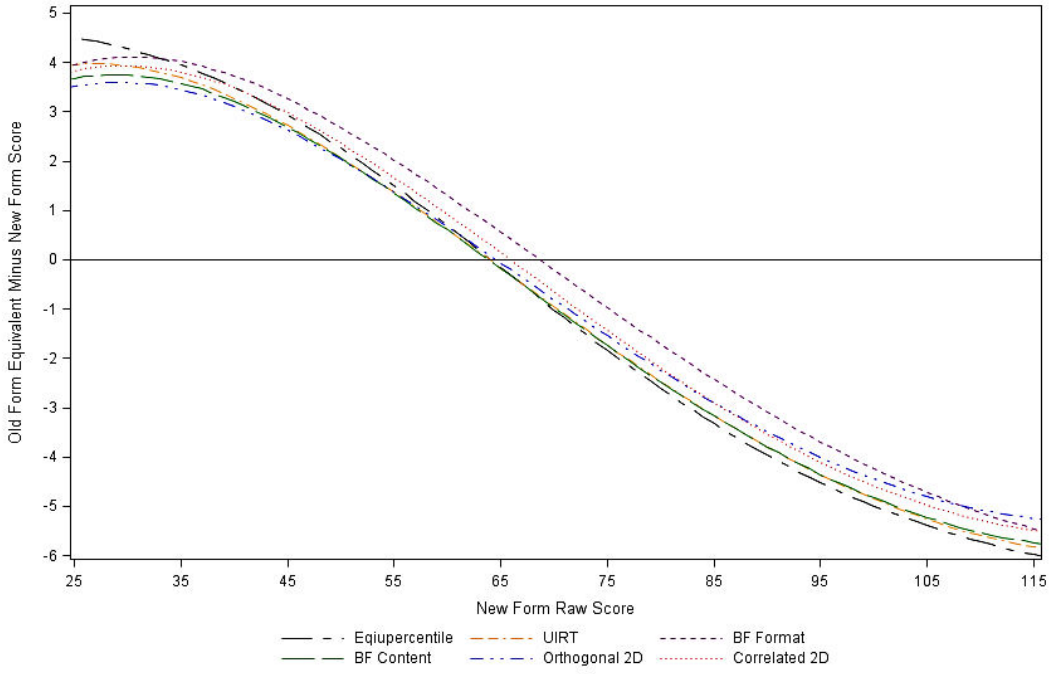


Figure 15. Equating relationships for English raw scores using all methods with the Matched Samples datasets.

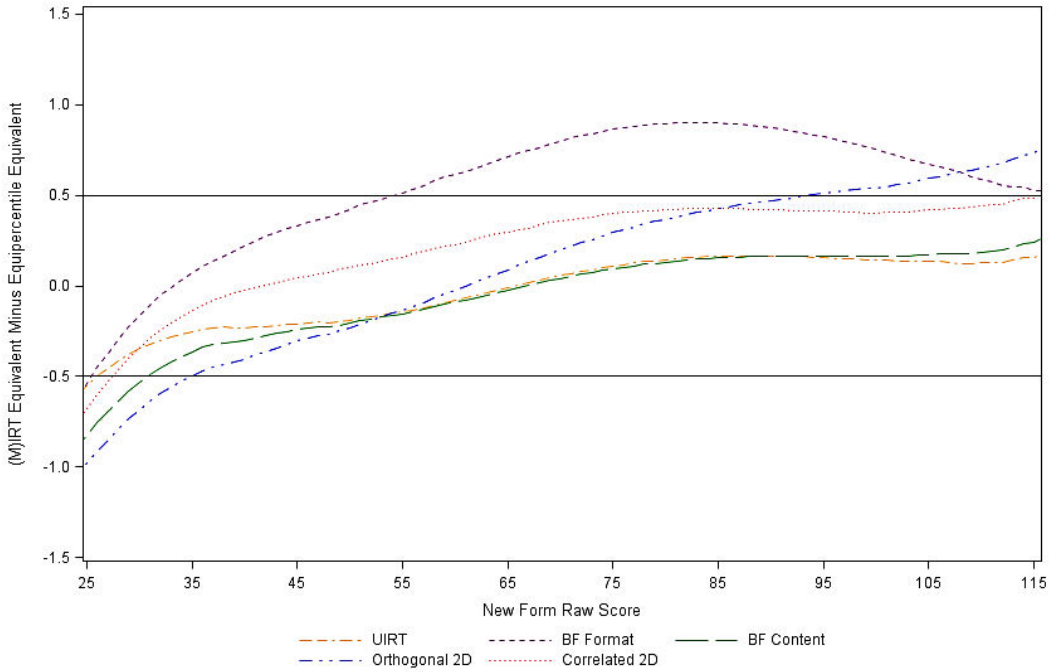


Figure 16. Differences between (M)IRT equivalents and equipercentile equivalents for English raw scores for the Matched Samples datasets.

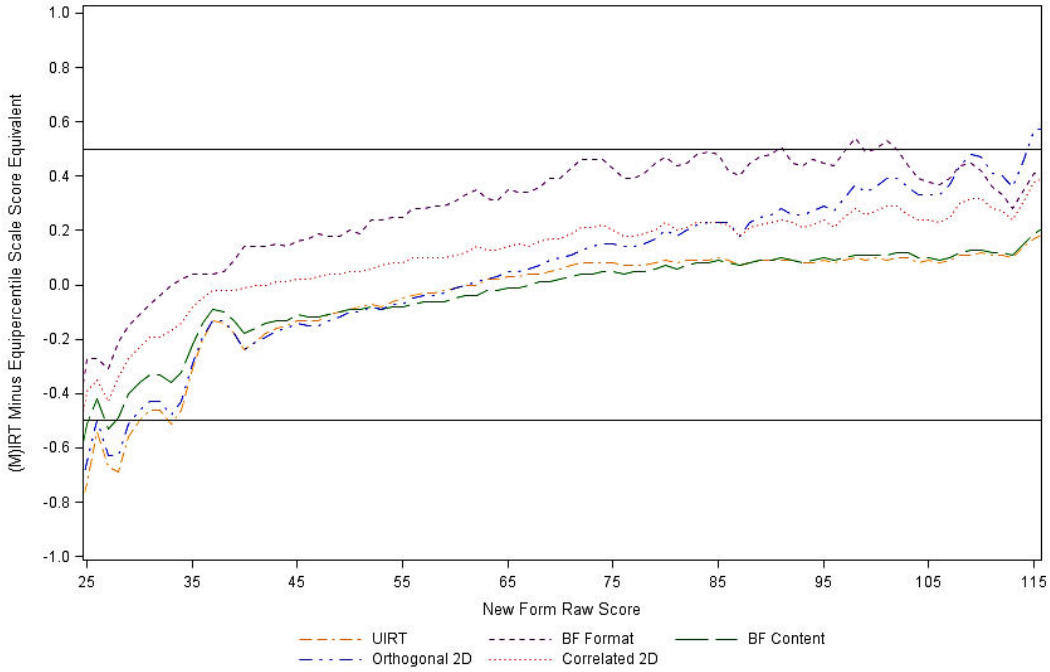


Figure 17. Differences between (M)IRT equivalents and equipercentile equivalents for English scale scores for the Matched Samples datasets.