Center for Advanced Studies in
Measurement and Assessment

CASMA Monograph

Number 2.5

Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 5)

Michael J. Kolen
Won-Chan Lee
(Editors)

May, 2018

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 5

This monograph, *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 5*), continues the work presented in Volumes 1-4 (Kolen & Lee, 2011, 2012, 2014, 2016). As stated in the Preface of the first volume,

> Beginning in 2007 … , 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.

Volume 5 contains 8 chapters. Chapter 1 provides an overview. In addition, it highlights some of the methodological issues encountered and some of the major findings. Chapters 2 through 8 address research primarily on psychometric methods for mixed-format exams.

We thank, CASMA Psychometrician Jaime Malatesta; past graduate students Shichao Wang, Mengyao Zhang, Yujin Kang, and Kyung Yong Kim; and current graduate students Stella Kim, Huan Liu, and Jiwon Choi for their work. We also thank Stella Kim for her effort in producing this volume.

We thank Robert L. Brennan who provided us with guidance where needed, and, who, as the founding director of CASMA, provided us with a home to conduct this work. Thanks to Anne Wilson for her 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 and Amy Hendrickson.

Michael J. Kolen

Won-Chan Lee

May, 2018

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.

Kolen, M. J., & Lee, W. (Eds.). (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3). (CASMA Monograph Number 2.3)*. Iowa City, IA: CASMA, The University of Iowa.

Kolen, M. J., & Lee, W. (Eds.). (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 4). (CASMA Monograph Number 2.4)*. Iowa City, IA: CASMA, The University of Iowa.

Contents

Chapter 1: Introduction and Overview for Volume 5

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

Abstract

This chapter provides an overview of this volume. It provides a brief description of the research questions, designs, and findings from each chapter. Where relevant, the findings from this volume are related to findings from the earlier volumes. The chapter concludes with a brief discussion.

Introduction and Overview for Volume 5

The research described in Volume 5 is closely related to research conducted in four previous volumes (Kolen & Lee, 2011, 2012, 2014, 2016). This chapter provides an overview of Volume 5 and 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. This chapter begins with a brief description of the research questions, designs, and findings from each chapter. Where relevant, the findings from Volume 5 are related to findings from the earlier volumes. The chapter concludes with a brief discussion.

Research Summary

The overall focus of this volume is on psychometrics for mixed-format tests, which are tests that contain both multiple-choice (MC) and free response (FR) items. Chapters 2 describes item response theory (IRT) approaches to evaluating psychometric properties of scores on mixed-format tests. Chapter 3 examines whether FR items built to same task models can be used as common items for equating purposes. Chapter 4 examines the minimum sample sizes needed for equipercentile equating under the random groups design. Chapter 5 evaluates a simple structure multidimensional IRT (MIRT) model for equating mixed-format tests. Chapter 6 examines the use of strong true score model equating with the common item nonequivalent groups design. Chapter 7 explores the dimensionality of scores for mixed-format tests. Chapter 8 investigates MIRT linking methods for the bifactor model. As suggested by these studies taken as a whole as well as studies in earlier volumes, psychometric analyses for mixed-format tests often require complex extensions of methods used with single-format tests.

Chapter 2

In Chapter 2, psychometric properties of raw and scale scores for mixed-format tests are evaluated under three IRT frameworks: unidimensional IRT, bifactor MIRT, and simple structure MIRT. Procedures are described for estimating conditional standard errors of measurement, classification consistency and accuracy, and reliability under each of the three IRT frameworks. Three AP exams with different levels of multidimensionality were used in the analysis. Results were similar across the three IRT frameworks, with the two MIRT models producing results that were more similar to each other than the unidimensional IRT framework..

Chapter 3

Chapter 3 focuses on the use of task models and equating, following up on a study by Malatesta and Liu (2016) that appeared in Volume 4. FR items on certain AP examinations are developed using task models. FR items on alternate forms of these tests are built to the same task model, and items built to the same task model share common characteristics. This study addressed the following question: Can items built from the same task model and appearing on alternate forms be treated as common items in equating? Although items built to the same task model were, at times, similar in difficulty, there was no compelling evidence to suggest that task model-derived FR items can be used as common-items for equating.

Chapter 4

The focus of Chapter 4 is on sample size requirements for random groups equating for mixed-format tests. Data were simulated based on data from a large-scale test. Equating error was estimated for different sample sizes and using smoothed equipercentile equating. In addition, the effect of violating the random groups assumption on equating error was investigated. This study extends the studies from earlier volumes that investigated smoothing (Liu & Kolen, 2011) and group differences (Powers et al., 2011) in equating.

Chapter 5

A true-score equating procedure based on a simple structure MIRT model is proposed in Chapter 5. The equating results from the proposed procedure are compared to results for four other procedures using both real data and simulation. The proposed procedure was found to have less error than unidimensional IRT equating procedures when there was substantial multidimensionality in the data. The work in this chapter extends work on MIRT equating from previous volumes (Lee & Brossman, 2012; Lee & Lee, 2014; Peterson & Lee, 2014).

Chapter 6

In Chapter 6, the accuracy of true score equating using the 4-parameter beta compound-binomial strong true-score model and its simplifications is investigated using real data and simulation. In this study, a 2-parameter beta simplification was generally most accurate of the methods investigated.

Chapter 7

The focus of Chapter 7 is on evaluating methods for assessing the dimensionality of mixed-format tests. This study reviews and compares an item-level exploratory factor analysis

procedure and a MIRT cluster analysis procedure for assessing dimensionality. Mixed-format test data are used in this comparison. The results suggest that there are inconsistencies in the number of dimensions and the clustering solution for the two methods. In addition, the results show how dimensionality assessment results are associated with the subject area, form, and sample size factors.

Chapter 8

In Chapter 8, a detailed description is provided of separate and concurrent calibration methods for scale transformation for the bifactor MIRT model. The two scale linking procedures are compared using a simulation study. In general, the concurrent calibration method was found to provide less linking error than the separate calibration method, demonstrating better recovery of the item parameters, test characteristic surface, and expected observed-score distribution.

Discussion and Conclusions

This volume along with Volumes 1, 2, 3, and 4 address many of the important psychometric issues associated with mixed-format tests. This work also reflects the use of a variety of different approaches to evaluating psychometric 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.

Kolen, M. J., & Lee, W. (Eds.). (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 3).* (CASMA Monograph Number 2.3). Iowa City, IA: CASMA, The University of Iowa.

Kolen, M. J., & Lee, W. (Eds.). (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 4).* (CASMA Monograph Number 2.4). Iowa City, IA: CASMA, The University of Iowa.

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) (pp. 115-142). Iowa City, IA: CASMA, The University of Iowa.

Lee, G., & Lee, W. (2014). A comparison of unidimensional IRT and Bi-factor multidimensional IRT 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) (pp. 201-233). Iowa City, IA: CASMA, The University of Iowa.

Liu, C., & Kolen, M. J. (2011). Evaluating smoothing in equipercentile equating using fixed smoothing parameters. 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. 213-236). Iowa City, IA: CASMA, The University of Iowa.

Malatesta, J., & Liu, H. (2016). Evaluating the interchangeability of free-response items developed from task models. 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. 77-111). Iowa City, IA: CASMA, The University of Iowa.

Peterson, J., & Lee, W. (2014). Multidimensional item response theory observed score 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 3).* (CASMA Monograph No. 2.3) (pp. 235-293). Iowa City, IA: CASMA, The University of Iowa.

Powers, S. J., Hagge, S. L., Wang, W., He, Y., Liu, C., & Kolen, M. J. (2011). Effects of group differences on mixed-format equating. 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. 51-73). Iowa City, IA: CASMA, The University of Iowa.

Chapter 2: IRT Approaches to Evaluating Psychometric Properties of Scores on Mixed-Format Tests

Won-Chan Lee, Stella Y. Kim, Jiwon Choi, and Yujin Kang
The University of Iowa, Iowa City, IA

Abstract

This paper considers psychometric properties of composite raw scores and transformed scale scores on mixed-format tests that consist of a mixture of multiple-choice and free-response items. Test scores on several mixed-format tests are evaluated with respect to conditional and overall standard errors of measurement, score reliability, and classification consistency and accuracy under three item response theory (IRT) frameworks: unidimensional IRT (UIRT), simple structure multidimensional IRT (SS-MIRT), and bifactor multidimensional IRT (BF-MIRT) models. Illustrative examples are presented using data from three mixed-format exams with various levels of format effects. In general, the two MIRT models produced similar results, while the UIRT model resulted in consistently lower estimates of reliability and classification consistency/accuracy indices compared to the MIRT models.

IRT Approaches to Analyzing Scores on Mixed-Format Tests

Psychometric properties of test scores often are evaluated to produce results that can be used to support and inform the use and interpretation of the scores (Kolen & Lee, 2011). This paper considers psychometric properties of composite raw scores and transformed scale scores on mixed-format tests that consist of a mixture of multiple-choice (MC) and free-response (FR) items. Scores on several mixed-format tests are evaluated with respect to conditional and overall standard error of measurement (SEM), score reliability, and classification consistency and accuracy under three item response theory (IRT) frameworks: unidimensional IRT (UIRT), bifactor multidimensional IRT (BF-MIRT), and simple structure multidimensional IRT (SS-MIRT) models. Although, strictly speaking, a *framework* differs from a *model* in the sense that different models can be considered within each framework, in this paper the terms are used interchangeably unless otherwise noted.

This paper begins with a general survey of the different item formats, followed by a discussion of psychometric issues associated with mixed-format tests, including dimensionality and score scales. Then the three IRT frameworks are presented, followed by a detailed description of the procedures for estimating various psychometric properties of scores under the three IRT frameworks. Finally, real data examples are illustrated and the paper concludes with a summary and discussion.

Overview of Mixed-Format Tests

According to the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), test item formats should be defined in the test specifications. Identification of appropriate item formats is typically driven by various considerations such as content specifications, testing time, item-writing resources, reliability, and scoring (Schmeiser & Welch, 2006). Among the many different types of items available, multiple-choice (MC) items are widely used because they tend to be expedient to score, have good domain coverage, and produce highly reliable scores. By contrast, free-response (FR) items are considered to be capable of tapping higher-order thinking skills and eliciting complex cognitive processes. Although FR items can take a variety of forms such as fill-in-the blank, short answer, extended response, essays, and computation, to name a few, such distinctions are not considered in this paper. Rather, the primary interest is whether different item formats are designed to assess different aspects of the constructs intended to be measured. For the sake of argument, only MC and FR are considered in this paper, although

the discussion and methodology presented here can easily be generalized to cases with more than two item types.

Given the advantages and disadvantages of different item formats, there has been an abiding interest in mixing different item types in a single test form to capitalize on the benefits of each item type for the purpose of gathering richer information about examinees' test performance. Using a mix of MC and FR items in a mixed-format test might afford the potential for coverage of a wide range of content domains with adequate reliability, and at the same time provoking deeper cognitive processes. However, combining MC and FR items in a test lends itself to many psychometric complexities. One such complexity is dimensionality.

Dimensionality

The current literature seems to be inconclusive as to whether MC and FR formats measure the same construct. A number of studies have demonstrated that MC and FR items measure essentially the same construct and that FR items fail to provide additional information beyond what can be obtained via MC items (Bennett, Rock, & Wang, 1991; Bridgeman, 1992; Lukhele, Thissen, & Wainer, 1994). However, in the study by Bridgeman (1992), substantial format effects were found at the item level, although total test scores in the FR and MC formats were comparable. The study conducted by Kennedy and Walstad (1997) revealed that examinees performed better on one item type than on the other, which resulted in a statistically significant number of misclassifications. Zhang, Kolen, and Lee (2014) evaluated dimensional structure for several mixed-format exams from different subject areas, and some format effects were identified in one of the exams. In summarizing research on the construct equivalence of MC and FR items, Traub (1993) concluded that the evidence is mixed. A meta-analysis by Rodriguez (2003) made a similar conclusion that the evidence for construct equivalence is inconclusive.

Different item formats are included in a test when the test developer believes that they can measure different and important constructs in the test domain (Hendrickson, Patterson, & Ewing, 2010). The confirmatory belief from the test development perspective can, in part, be supported by some exploratory analyses to provide further evidence that the MC and FR items measure different traits. In such cases where MC and FR items are judged to measure somewhat distinct traits, using UIRT models for analyzing the data may not be appropriate; instead, multidimensional models could be considered (Cao, 2008; Kolen & Lee, 2011; Lee & Lee, 2016; Peterson & Lee, 2014; Yao & Boughton, 2009). In this paper, UIRT and two alternative MIRT models, BF-MIRT and SS-MIRT, are considered and compared.

Composite Scores and Transformed Scales Scores

Interpretation and use of a test depends heavily on the types of scores that are associated with examinees' performance on the test. When the items in different formats are presumed to measure essentially the same or similar traits, combining the scores from each format to form a composite total score would be easily justifiable (Wainer & Thissen, 1993). However, when different item formats are designed to tap substantially different content or cognitive domains, the justification for using a composite score would become much more complex, and reporting separate component scores can be considered (Rodriguez, 2003). Various methodological and conceptual issues on subscore reporting have been well documented (Brennan, 2011; Haberman, 2008; Haberman, Sinharay, & Puhan, 2009; Sinharay, Puhan, & Haberman, 2011). However, for most large-scale educational testing programs that employ mixed-format tests, the primary score scale of interest is likely to be the composite total score.

Two additional complexities exist regarding reported scores. First, the scores on different item formats can be weighted, resulting in a weighted composite score. Since the MC and FR items typically have different numbers of score points, weights are often identified in relation to the contribution of each item to the total in terms of the number of score points. A variety of weighting schemes can be considered—for example, see Kolen (2006) and Wainer and Thissen (1993). The weights for different item formats can be either integers or non-integers. Assuming there are two item formats, MC and FR, let the composite score, *Y*, be the rounded integer score defined as

$$Y = int(w_{MC}X_{MC} + w_{FR}X_{FR}), \quad (1)$$

where X_{MC} and X_{FR} represent the MC and FR section scores, respectively, which are computed by summing the item scores in each section; and w_{MC} and w_{FR} are weights associated with MC and FR sections, respectively. Note that the $int()$ function in Equation 1 returns integer scores when non-integer weights are employed.

Another complexity involves transforming the composite raw scores to scale scores for reporting purposes. More often than not, the transformation is nonlinear, which makes many psychometric issues much more complex. Types of nonlinearly transformed scale scores include normalized scores with a particular mean and standard deviation, percentile ranks, variance stabilizing arcsine transformation, etc. (Kolen, 2006). In notation, a scale score *S* is a nonlinear transformation of a composite raw score *Y*, such that $S = t(Y)$, where t indicates the transformation function. Scale scores are typically rounded to integers when

reported to examinees. One of the goals of this paper is to estimate psychometric properties of transformed scale scores as well as the weighted composite raw scores on mixed-format tests under each of the three IRT frameworks that are discussed next.

IRT Frameworks for Mixed-Format Tests

The three IRT frameworks considered in this paper are UIRT, BF-MIRT, and SS-MIRT models, which provide different ways of dealing with potential format effects. UIRT models have been predominantly used in educational testing including mixed-format tests, even if the assumption of unidimensionality is often untenable. Two relatively simple MIRT models, BF- and SS-MIRT, have been gaining popularity in recent years and have been increasingly applied in psychological and educational measurement (Cai, Yang, & Hansen, 2011; DeMars, 2006, 2013; Gibbons & Hedeker, 1992; Gibbons et al., 2007; Kolen, Wang, & Lee, 2012; Rijmen, 2010). In particular, both models have been considered in test equating with mixed-format tests (Lee & Brossman, 2012; Lee & Lee, 2016; Peterson & Lee, 2014). The BF-MIRT and SS-MIRT models might be viewed as special cases of a more general two-tier item factor analysis model proposed by Cai (2010). Compared to the full MIRT model, which is not considered in this paper, the BF-MIRT and SS-MIRT models have a principal advantage of computational efficiency involving a much smaller number of integral evaluations (Cai, 2010).

Model Specifications

Under the UIRT framework, it is assumed that both MC and FR items are designed to measure the same general trait, θ_G. Any potential extra dimensions due to different item formats are considered trivial and thus are not modeled. In identifying a UIRT model, the latent distribution of θ_G is typically set to have a mean of zero and standard deviation of one.

Under the SS-MIRT framework, each item is associated with a trait corresponding to each specific item type, θ_{MC} or θ_{FR}, for the MC and FR items, respectively. The two latent traits associated with the two item types are allowed to be correlated. In model estimation, each of the θ_{MC} and θ_{FR} latent distributions is assumed to have a mean of zero and standard deviation of one with a correlation between the two traits. The underlying assumption of the SS-MIRT model is that each item type, by design, is intended to measure different, but related constructs. The SS-MIRT model has unique advantages that (a) a subset of items with the same item type is modeled by a UIRT model and (b) because each item loads on only one trait, interpretation is straightforward.

By contrast, under the BF-MIRT framework, all items in the test are presumed to measure a general trait, θ_G; in addition to θ_G, each MC and FR item is assumed to measure at most one additional trait that is specific to the item type, θ_{MC} or θ_{FR}. The three latent traits are typically assumed to be uncorrelated, and each of the three traits is set to have a mean of zero and standard deviation of one in calibration. Unlike the SS-MIRT model, the two specific traits, θ_{MC} and θ_{FR}, for the BF-MIRT model represent residual factors after controlling for the general trait. Depending upon the contexts, the general trait can sometimes be the focus of interpretation and whatever is left in the specific components are considered to be nuisance factors (DeMars, 2013). In some other applications, such as the present consideration of a mixed-format test, attention is given to not only the general trait, but the residual factors after controlling for the general factor. For example, the test developer may be interested in assessing general writing ability as well as the factual recollection of basic writing rules represented by MC items and the logical expression in writing demonstrated by FR items.

Figure 1 depicts diagrams for the three IRT frameworks. In each diagram in the figure, there are seven items, four MC and three FR items, with θ_G, θ_{MC}, and θ_{FR} representing the general, MC-related, and FR-related traits, respectively. Note that the number of latent traits specified is one, two, and three for the UIRT, SS-MIRT, and BF-MIRT models, respectively, which plays an important role in identifying the latent-trait distributions when estimating some psychometric properties.

In sum, the following presents the key assumptions for each of the three IRT frameworks:

- **UIRT**: MC and FR items are intended to measure essentially the same single trait, and any unintended traits due to item formats are negligible.
- **SS-MIRT**: Each item format is intended to measure its own trait, and the two traits measured by the MC and FR items, respectively, are different, yet can be correlated.
- **BF-MIRT**: MC and FR items are intended to measure the same general trait, and each item format is intended to measure an additional trait, which is uncorrelated with the general or the other item format.

Psychometric Properties of Composite Raw and Scale Scores

The psychometric properties of interest that are considered in this paper include (a) conditional SEM (CSEM), (b) score reliability, and (c) classification consistency and accuracy. Methods for estimating CSEMs and reliability under a UIRT framework have been

proposed and studied in the literature (Kolen, Zeng, & Hanson, 1996; Lee, Brennan, & Kolen, 2000; Wang, Kolen, & Harris, 2000). Classification consistency and accuracy has also been discussed in the literature under the UIRT framework (Huynh, 1990; Lee, 2010; Schulz, Kolen, & Nicewander, 1999). Knupp (2009) and LaFond (2014) employed SS-MIRT and BF-MIRT, respectively, to examine classification indices. Kim and Lee (2016) considered classification consistency and accuracy for mixed-format exams under the three IRT frameworks. Building upon the previous research, one of the goals of the present study is to extend the existing IRT procedures for estimating these psychometric properties for composite raw scores and nonlinearly transformed scale scores using BF-MIRT and SS-MIRT as psychometric frameworks. In particular, a few alternative estimation methods with respect to the latent-trait variables are considered and compared in the real data examples. Of particular interest are the psychometric property results conditional on (multidimensional) latent traits.

Theoretical Framework

A useful psychometric model would include all of the following three components in the model: observed score, latent trait or true score, and measurement error. The fundamental equation that underlies all three IRT frameworks expresses the marginal composite raw-score distribution as:

$$\Pr(Y = y) = \int_{\boldsymbol{\theta}} \Pr(Y = y|\boldsymbol{\theta})\, g\boldsymbol{\theta} d\boldsymbol{\theta}, \tag{2}$$

where Y is defined in Equation 1, $\Pr(Y = y|\boldsymbol{\theta})$ is the conditional composite raw-score distribution, and the conditioning variable $\boldsymbol{\theta}$ depends on the IRT model—for example, for UIRT $\boldsymbol{\theta}$ is a scalar θ_G, for SS-MIRT $\boldsymbol{\theta} = \{\theta_{MC}, \theta_{FR}\}$, and for BF-MIRT $\boldsymbol{\theta} = \{\theta_G, \theta_{MC}, \theta_{FR}\}$. The conditional distribution, $\Pr(Y = y|\boldsymbol{\theta})$, in Equation 2 represents the score variability for an examinee with $\boldsymbol{\theta}$ over repeated measurement (i.e., measurement error). An efficient formula for calculating the conditional distribution with any set of weights for each item format, either integer or non-integer, involves separating the section scores, which is given by

$$\Pr(Y = y|\boldsymbol{\theta}) = \sum_{Y = int(w_{MC}X_{MC} + w_{FR}X_{FR})} \Pr(X_{MC} = x_{MC}|\boldsymbol{\theta}) \Pr(X_{FR} = x_{FR}|\boldsymbol{\theta}), \tag{3}$$

where the multiplication of the two section-score probabilities is possible due to the local independence assumption and the summation is taken over all possible combinations of section scores that lead to the same rounded composite raw score. The conditional

distribution for each section can be computed using recursive formulas (Hanson, 1994; Lord & Wingersky, 1984; Thissen, Pommerich, Billeaud, & Williams, 1995). If the section weights are all integers, the separation by section scores in Equation 3 is not necessary, and in that case, the recursive formula can be directly applied to the composite scores.

CSEMs and Reliability

A similar presentation of formulas under the UIRT framework can be found in Kolen et al. (1996) and Kolen and Lee (2011). The mean of the conditional distribution, $\Pr(Y = y|\boldsymbol{\theta})$, is true composite raw score, which is given by

$$\tau_{\boldsymbol{\theta}} = \sum y \cdot \Pr(Y = y|\boldsymbol{\theta}), \tag{4}$$

where the summation is taken over the entire range of integer values of *Y*. The CSEM for composite raw scores at $\boldsymbol{\theta}$ is the standard deviation of $\Pr(Y = y|\boldsymbol{\theta})$:

$$\sigma_{Y|\boldsymbol{\theta}} = \sqrt{\sum (y - \tau_{\boldsymbol{\theta}})^2 \cdot \Pr(Y = y|\boldsymbol{\theta})} = \sqrt{\sum y^2 \cdot \Pr(Y = y|\boldsymbol{\theta}) - \tau_{\boldsymbol{\theta}}^2}, \tag{5}$$

where the summations are over the range of integer *Y*. The overall error variance for the population can be computed by integrating the squared CSEMs in Equation 5, which can then be used to obtain reliability for composite raw scores as:

$$\rho_{YY'} = 1 - \frac{\int_{\boldsymbol{\theta}} \sigma_{Y|\boldsymbol{\theta}}^2 g\boldsymbol{\theta} d\boldsymbol{\theta}}{\sigma_Y^2}, \tag{6}$$

where σ_Y^2 is the variance of the composite raw scores based on the model [i.e., variance of $\Pr(Y = y)$ in Equation 2].

The results for scale scores, $S = t(Y)$, can be obtained in an analogous manner. The conditional scale-score distribution is obtained using the fact that $\Pr[t(Y)|\boldsymbol{\theta}] = \Pr(Y|\boldsymbol{\theta})$, where $\Pr(Y|\boldsymbol{\theta})$ is defined in Equation 3. Likewise, for the marginal distribution, $\Pr[t(Y)] = \Pr(Y)$. Replacing *y* with $t(Y)$ in Equation 4 gives true scale score $\xi_{\boldsymbol{\theta}}$. The CSEMs for scale scores, $\sigma_{S|\boldsymbol{\theta}}$, are obtained by substituting $t(Y)$ for *y* and $\xi_{\boldsymbol{\theta}}$ for $\tau_{\boldsymbol{\theta}}$ in Equation 5. Finally, the IRT model-based reliability for scale scores can be defined in the same way as Equation 6 using *S* instead of *Y*.

Classification Consistency and Accuracy

When test scores are used to determine the performance levels of examinees with respect to a set of criteria or cut scores, the conventional reliability definition presented in Equation 6 may not provide adequate information. Dependability of classifications is often

discussed in terms of classification consistency and accuracy (Lee, 2010; Livingston & Lewis, 1995). The formulation of classification indices presented here is similar to Lee (2010) and Kim and Lee (2016). Suppose the classifications of examinees for *H* mutually exclusive performance levels are operated based on a set of cut scores, $c_1, c_2, \dots, c_{H-1}$, on the scale of the integer composite raw score, *Y*. Examinees with an observed composite raw score between c_{h-1} and c_h are classified into level L_h $(h = 1, 2, \dots, H)$, with $Y < c_1$ belonging to L_1 and $Y \geq c_{H-1}$ belonging to L_H. Let Y_1 and Y_2 denote the composite raw scores from two independent administrations of the same test such that the conditional distribution defined in Equation 3 is the same for Y, Y_1, and Y_2.

The classification consistency index conditional on $\boldsymbol{\theta}$ can be defined as:

$$\phi_{\boldsymbol{\theta}} = \sum_{h=1}^{H} \Pr(Y_1 \in L_h, Y_2 \in L_h | \boldsymbol{\theta}) = \sum_{h=1}^{H} [\Pr(Y \in L_h | \boldsymbol{\theta})]^2. \quad (7)$$

The conditional probability for each level, $\Pr(Y \in L_h | \boldsymbol{\theta})$, in Equation 7 can be calculated as

$$\Pr(Y \in L_h | \boldsymbol{\theta}) = \sum_{y=c_{(h-1)}}^{c_h - 1} \Pr(Y = y | \boldsymbol{\theta}), \quad h = 1, 2, \dots, H, \quad (8)$$

where $c_0 = min(Y)$ when $h = 1$ and $c_H - 1 = max(Y)$ when $h = H$; and the conditional distribution, $\Pr(Y = y | \boldsymbol{\theta})$, is from Equation 3. The overall classification consistency index for the population is

$$\phi = \int_{\boldsymbol{\theta}} \phi_{\boldsymbol{\theta}} g \boldsymbol{\theta} d \boldsymbol{\theta}. \quad (9)$$

Classification accuracy describes to what extent the classifications based on observed scores match with those based on true scores. The true performance level of each examinee can be determined by comparing $\tau_{\boldsymbol{\theta}}$ in Equation 4 with the cut scores. Suppose the true score for an examinee with $\boldsymbol{\theta}$ falls in the J^{th} performance level, $\tau_{\boldsymbol{\theta}} \in L_J$. The probability of an accurate classification for $\boldsymbol{\theta}$ is simply

$$\gamma_{\boldsymbol{\theta}} = \Pr\big(Y \in L_J | \boldsymbol{\theta}\big). \quad (10)$$

That is, an accurate classification occurs when an examinee is classified in a performance level, which is the same as the examinee's true performance level. The overall classification accuracy index for the entire population is

$$\gamma = \int_{\boldsymbol{\theta}} \gamma_{\boldsymbol{\theta}} g\boldsymbol{\theta} d\boldsymbol{\theta}. \quad (11)$$

Classification accuracy can also be described with respect to the false positive and negative error rates. A false positive error occurs when the assigned level based on the observed score is higher than the true level, whereas a false negative error occurs when the observed level is lower than the true level. The conditional positive and negative error rates, respectively, are:

$$\gamma_{\boldsymbol{\theta}}^{+} = \sum_{h=J+1}^{H} \Pr(Y \in L_h | \boldsymbol{\theta}), \quad (12)$$

where $\gamma_{\boldsymbol{\theta}}^{+} = 0$ when $J = H$; and

$$\gamma_{\boldsymbol{\theta}}^{-} = \sum_{h=1}^{J-1} \Pr(Y \in L_h | \boldsymbol{\theta}), \quad (13)$$

where $\gamma_{\boldsymbol{\theta}}^{-} = 0$ when $J = 1$. The overall error rates, γ^{+} and γ^{-}, can be computed by integrating the conditional error rates over the $\boldsymbol{\theta}$ distribution.

The foregoing presentation of classification indices is pertinent to cases in which the cut scores are expressed on the composite raw-score scale. In some cases, cut scores may be placed on the metric of reported scale scores. One way for computing classification indices for scale scores is to use the same formulas previously presented but by replacing Y with $S = t(Y)$. For example, the conditional distribution of scale scores can be obtained by substituting $t(Y)$ for Y in Equation 3, which, in turn, is used in all the subsequent steps for computing classification indices. An alternative way is to identify the composite raw-score points that correspond to the scale-score cut scores and then use the formulas presented here as they are. In case a scale-score cut-score point has multiple corresponding composite raw-score points, the lowest of the corresponding points should be selected. These two approaches will yield the same result.

Estimation Methods

All of the equations in previous sections are expressed in terms of parameters. Estimation of psychometric properties begins with estimating item and latent-trait parameters for a given IRT model. There are at least three methods that can be used in lieu of the latent-trait parameter $\boldsymbol{\theta}$ based each on: (a) a quadrature distribution (D-method), (b) individual latent-trait estimates $\widehat{\boldsymbol{\theta}}$ (P-method), and (c) Monte Carlo simulation (M-method).

D-method. A discrete quadrature distribution for $\boldsymbol{\theta}$ is specified for the D-method using either a standard normal or a posterior distribution. In this study, the standard normal distribution is employed for all three IRT frameworks. For a UIRT model, a set of quadrature points (say, $n_q = 41$) and associated density (i.e., weights w_q such that $\sum w_q = 1$) for a univariate standard normal distribution, UVN(0,1), is identified. For the SS-MIRT model with two item formats, a set of pairs of quadrature points and weights are obtained from a bivariate standard normal distribution, BVN($\mathbf{0}, \mathbf{1}, \hat{\rho}$) with an estimated latent-trait correlation, $\hat{\rho}$. With $n_q = 41$ for each dimension, there are $41 \times 41 = 1{,}681$ pairs of theta values that need to be evaluated. An estimate of the latent-trait correlation ($\hat{\rho}$), can be obtained from a calibration program such as flexMIRT (Cai, 2013); alternatively, a classical disattenuated correlation could be used as a proxy when a test is relatively long. Finally, the BF-MIRT model requires three dimensions from a trivariate standard normal distribution, TVN($\mathbf{0}, \mathbf{1}, \hat{\rho}$), with $\hat{\rho} = 0$ correlations, resulting in $41^3 = 68{,}921$ combinations of theta values that are to be evaluated.

For this method, when conditional results are integrated over the $\boldsymbol{\theta}$ distribution to obtain the overall statistics (i.e., Equations 6, 9, and 11), the integrals are replaced with summations. For example, the overall classification consistency index ϕ under the SS-MIRT model is computed as: $\hat{\phi} = \sum_q \hat{\phi}_{\boldsymbol{\theta}} w_q$, where the summation is taken over all 1,681 pairs of theta values.

P-method. The parameter $\boldsymbol{\theta}$ in the equations is replaced by $\widehat{\boldsymbol{\theta}}$ for the P-method. For each person, the latent trait(s) can be estimated using one of the IRT latent-trait estimation methods such as the maximum likelihood estimate, expected a posteriori estimate, and Warm's weighted likelihood estimate (Bock & Mislevy, 1982; Lord, 1980; Warm, 1989). Conditional results are computed for each person using $\widehat{\boldsymbol{\theta}}$, and then the overall results are obtained by taking the average of the conditional results over the number of persons in the entire group. For example, $\hat{\phi} = \sum \hat{\phi}_{\widehat{\boldsymbol{\theta}}} / N$, where N is the number of persons in the data.

M-method. For the M-method, a large sample of random deviates is drawn from a specified $\boldsymbol{\theta}$ distribution for a given IRT model. For example, a large number (e.g., $N = 10{,}000$) of pairs of correlated theta values are randomly drawn from BVN($\mathbf{0}, \mathbf{1}, \hat{\rho}$) for the SS-MIRT model. For each of the *N* simulees, conditional results are computed. Similar to the P-method, the overall results are computed by taking the average of the conditional results over the number of simulees, *N*.

Both the D-method and M-method are considered in the present study. The P-method is not considered because the results might be affected by the choice of a theta estimation method; in practice, either the D- or M-method might be easier to implement. Some advantages of the M-method compared to the D-method are discussed later.

Real Data Examples

Multiple real data examples from the Advance Placement (AP) Examinations are used to illustrate the procedures for estimating psychometric properties for mixed-format tests. Because the data were manipulated in various ways for illustrative purposes, the results reported in this study should not be directly applied to operational AP examinations.

Data

Three sets of mixed-format data were obtained from the AP Examinations. Data from one form of each of the following AP exams were used: Spanish Literature, English Language and Composition, and US History. These exams have varied characteristics in terms of subject areas, test length, composition of MC and FR items, and degree of format effects. Each exam has a conversion table in which the composite raw scores convert to normalized scale scores ranging from 0 to 70. Psychometric properties are estimated for both the composite raw and scale scores.

Table 1 summarizes the characteristics of the three exams, including the first four moments of the weighted composite raw scores, numbers of MC and FR items, sample size, estimated latent trait correlation between the two item formats, and cut scores that were used for computing classification indices. Note that integer section weights were used for all three exams. Of particular interest is the magnitude of the estimated latent-trait correlations, $\hat{\rho}_{\theta_{MC}\theta_{FR}}$, which is an indicator of the extent to which the two item formats measure different constructs. The estimated classical disattenuated correlations, $\hat{\rho}_{T_{MC}T_{FR}}$, are also provided, for which coefficient alpha was used as a reliability estimate for scores on each item-format section. It appears that the IRT-based latent-trait correlations and the classical disattenuated correlations are very similar to each other, which is often the case for long tests. More importantly, the Spanish exam has the lowest correlation, indicating substantial multidimensionality due to format effects; in contrast, the History exam has the largest correlation, suggesting that the data might be essentially unidimensional. The English exam lies in between, with a moderate level of multidimensionality.

The MC and FR items in the AP exams are designed to measure different skills or knowledge in a particular subject area. For example, for the English Language and

Composition exam, College Board (2014) states that the MC items "… test students' comprehension of the literal meaning of the text, their ability to infer the writer's intended meaning from the formal features of the text…;" whereas one of the three essay prompts "… requires students to address an issue by synthesizing information from multiple texts." The specific definitions of what each item type can measure in the test design provides a strong support for exploring the applicability of the two MIRT models considered in the present paper.

Analysis and Model Fit

Item calibrations were conducted using flexMIRT (Cai, 2013), which has the full capability of estimating item and latent-trait parameters for all three IRT models considered in this study. For the dichotomously scored MC items, the three-parameter logistic (3PL) model was used for UIRT and SS-MIRT, while the bifactor 3PL model (Cai et al., 2011; DeMars, 2013) was used for BF-MIRT. The polytomously scored FR items were fit with the graded response (GR) model for UIRT and SS-MIRT, and with the bifactor version of GR model (Cai et al., 2011; DeMars, 2013) for BF-MIRT.

The degree of model fit of the three IRT models was assessed using several fit indices and results are summarized in Table 2. Test-level fit statistics examined include the Akaike Information Criterion (AIC; Akaike, 1981), Bayesian Information Criterion (BIC; Schwarz, 1978), and root mean square error of approximation (RMSEA; Steiger & Lind, 1980). Orlando and Thissen's (2000) $S - X^2$ item fit statistic was also considered. Results for all these fit indices were obtained from the output of flexMIRT. Taken all together, there seems to be a general tendency for the two MIRT models to provide somewhat better fit than UIRT for the Spanish and English exams, which are associated with relatively larger format effects.

In addition to the fit indices, the model-fitted observed score distributions are plotted against the actual frequency distributions in Figure 2. A good-fitting model would produce a fitted distribution that is smooth and closely follows the actual frequency distribution. All three models tend to follow the actual score distributions fairly well for all three exams. The three IRT models overlap significantly for the US History exam, indicating that they provide a similar level of model fit for the essentially unidimensional data. By contrast, for the most multidimensional exam (i.e., Spanish), the BF-MIRT model tends to produce a fitted distribution that is quite different from those based on the UIRT and SS-MIRT models. For the moderate multidimensional data (i.e., English), the BF-MIRT is still somewhat distinct from the other two models, but the difference is small. One interesting point is that the close proximity of the fitted distributions between the UIRT and SS-MIRT models doesn't

necessarily lead to similar results in terms of the psychometric properties. In fact, as discussed in the next section, the psychometric properties are more similar between the two MIRT models than the UIRT model.

For the D-method, computations for all of the psychometric properties were conducted using 41 quadrature points for each dimension. Several sample sizes were tried out for the M-method, and $N = 10{,}000$ was determined to be appropriate with respect to both the stability of results and computational tractability. All the results reported in this paper were obtained using the computer program MIRT-PP (Lee, 2015), which is available upon request.

Results for CSEMs and Reliability

The overall SEMs (i.e., the numerator of the fraction in Equation 6) and reliability estimates for composite raw scores are summarized in Table 3. A few observations from the table are worth mentioning. First, the D-method and the M-method produced almost congruent results across all three exams. When large discrepancies occur between the D- and M-methods in practice, increasing the number of simulees with the M-method usually helps. In the present study, using $N = 10{,}000$ seemed to work adequately. Second, the two MIRT models tend to yield results that are more similar to each other than to the results for the UIRT model. As mentioned previously, this result is somewhat contrary to the model-fit results based on the fitted observed-score distributions. The differences between the UIRT vs. the two MIRT models become much smaller for the US History exam, which may be viewed as being essentially unidimensional. Third, the reliability estimates for UIRT are always smaller than those for the two MIRT models—the opposite is true for the overall SEMs.

Table 4 contains the same set of results as Table 3 for scale scores. Similar to the composite raw-score results, the D- and M-methods show results that are highly comparable, and the results for the two MIRT models are closer to each other than UIRT. While the reliability estimates for scale scores are almost identical to those for the composite raw scores, the SEM values for scale scores are much smaller than those for the composite raw scores because the SEMs are expressed in the score-scale units under consideration.

The CSEMs for composite raw scores based on the D-method are plotted in Figure 3. In order to avoid displaying the CSEMs in the multidimensional $\boldsymbol{\theta}$ space, the CSEMs are plotted against the true composite raw score, $\tau_{\boldsymbol{\theta}}$ (see Equation 4). Because of the one-to-one relationship between the unidimensional θ and true score under the UIRT framework, the CSEMs evaluated at 41 quadrature points appear as a smooth line. By contrast, the CSEMs

under the two MIRT models tend to exhibit vertical scatter, which is due to the fact that there could be a number of pairs or combinations of theta quadrature points that are related to the same true composite raw score. Notice also that, unlike the UIRT and SS-MIRT models, the results for BF-MIRT do not show any white spaces between dots, which was ensured by the substantially large number of theta combinations (= 68,921) that were used with the BF-MIRT model.

Overall, across all exams and IRT models, the patterns of CSEMs for composite raw scores shown in Figure 3 tend to be concave in shape—namely, smaller error in both ends and larger error in the middle of the score scale. This is consistent with the findings from previous studies (e.g., Brennan & Lee, 1999; Lee et al., 2000). Some exceptions exist with the MIRT models, especially with the SS-MIRT model—there are a few true-score points in the middle score range where the CSEMs have substantial vertical variability and can be very low (e.g., true score of 80 for US History). The umbrella pattern of CSEMs is unique to the D-method. The greater variability of the SS-MIRT CSEMs at true score of 80 for US History indicates that there are many pairs of quadrature points resulting in true scores near 80, and some of the pairs with very low CSEMs are those that are not likely to observe in reality; for example, a pair of $\theta_{MC} = -3$ and $\theta_{FR} = +3$ for an examinee may not be a realistic combination. This illustrates a drawback of the D-method in that it is usually set up in such a way that all possible combinations of multivariate theta values are evaluated. Exclusion of some of the unrealistic combinations of theta values can be considered, but doing so requires considerable subjective decisions. This problem is greatly mitigated with the M-method as discussed later.

Figure 4 presents plots for scale-score CSEMs based on the D-method. The horizontal axis in each plot is true scale score, $\xi_{\boldsymbol{\theta}}$. The patterns of scale-score CSEMs appear quite different from those for the composite raw scores. As previous research suggests (Brennan & Lee, 1999; Lee et al., 2000), scale-score CSEMs have a strong tendency to follow the pattern of changes in the slope (i.e., first derivative), along the score scale, of the raw-to-scale score transformation function. For US History, for example, the scale-score CSEMs are quite small near true scale scores between 20 and 30. Although not reported here, the scale-score conversion function for US History, when plotted, is quite flat near the composite raw-score points that are converted to scale scores between 20 and 30. A steeper slope at a composite raw score leads to a larger scale-score CSEM, whereas a flatter slope leads to a smaller scale-score CSEM. In general, the overall patterns of scale-score CSEMs are similar for the three IRT frameworks, save for the vertical scatter with the MIRT models.

The CSEM results based on the M-method are provided in Figures 5 and 6 for composite raw scores and scale scores, respectively. It is apparent that the results for the two MIRT models show much less vertical variability compared to the results from the D-method. Taking the SS-MIRT model as an example, the principle reason for that is that, with the M-method, there is only a slim chance for extremely unusual pairs of theta values to be drawn from a bivariate distribution with a relatively high correlation. In this regard, the M-method might be preferred over the D-method when used in practice.

Oftentimes, it is practically desirable to report a single value of CSEM at each true-score point when a MIRT model is used. One way to do so would be to compute an arithmetic mean of the CSEMs for each true score; however, doing so necessitates creating some arbitrary intervals around each true score because the true scores resulting from applying Equation 4 are typically non-integers. Another way, which is probably much easier to implement, would be to fit a high-degree polynomial regression on the CSEMs. More precisely, the conditional error variances are fitted with a polynomial, for which the square root is then obtained. It would be much more sensible to use the CSEMs based on the M-method than the D-method when fitting a polynomial because, with the D-method, each pair or combination of theta values, including those unusual ones, contribute equally to the fitted polynomial. Figure 7 displays the polynomial-fitted CSEMs for both composite raw and scale scores based on the results from the M-method. A polynomial with degree 2 was fitted to the composite raw-score CSEMs for all exams; for scale scores, degree 6 was used for all exams. Notice in Figure 7 that the UIRT CSEMs are almost always larger than the CSEMs for the two MIRT models; however, the discrepancies become smaller with less multidimensional data, US History. Results for the two MIRT models are very similar for English and US History, while all three models tend to be quite different for Spanish, which is the most multidimensional exam.

Results for Classification Consistency and Accuracy

The estimated overall classification consistency and accuracy indices are presented in Table 5. With respect to the results for both $\hat{\phi}$ and $\hat{\gamma}$, (a) UIRT always has lower estimates than the two MIRT models, (b) the two MIRT models are closer to each other than to UIRT, and (c) the results based on the D- and M-methods are very similar. Note also that $\hat{\gamma}$ is always larger than $\hat{\phi}$, which has been found in previous research (e.g., Lee, Hanson, & Brennan, 2002). Results for the false positive and negative error rates indicate that the estimates based on the MIRT models are similar to each other, whereas UIRT estimates are

quite different from the two MIRT models. The false positive error rates tend to be larger than the false negative error rates for these particular examples. The relative importance of the two error rates typically depends on the purpose of and use of scores on a particular test.

Figures 8 and 9, respectively, display conditional classification consistency ($\hat{\phi}_{\boldsymbol{\theta}}$) and accuracy ($\hat{\gamma}_{\boldsymbol{\theta}}$) indices computed using the D-method. Similar to the CSEM plots, conditional classification results are plotted against true scores. The jagged patterns of conditional classification indices suggest that the degree of consistency and accuracy deteriorates as the true score comes close to one of the cut scores. That is, chances are higher that examinees whose true scores are near the cut scores are placed on a performance level that is either inaccurate or inconsistent if retested. Vertical scatter is again shown in the results for the two MIRT models.

Displayed in Figures 10 and 11, respectively, are conditional classification indices, $\hat{\phi}_{\boldsymbol{\theta}}$ and $\hat{\gamma}_{\boldsymbol{\theta}}$, based on the M-method. Compared to the results for the D-method, there is much less vertical scatter and all three IRT models produced results that are similar in terms of both pattern and magnitude.

Summary and Discussion

Using IRT for mixed-format tests requires a careful check for the assumption of unidimensionality. When there is evidence that format effects exist either by the test design or empirical analysis of dimensionality, the use of a UIRT model will lead to erroneous parameter estimates, which propagate in subsequent psychometric analyses resulting in inadequate use and inaccurate interpretation of test scores. This paper considers two alternative models, SS-MIRT and BF-MIRT, which take into account the format effects in their model specifications for estimating psychometric properties of scores on mixed-format tests.

Various psychometric properties are considered, including overall and conditional SEMs, score reliability, and overall and conditional classification consistency and accuracy indices for both weighted composite raw scores and transformed scale scores. The procedures for estimating these psychometric properties under each of the UIRT, SS-MIRT, and BF-MIRT frameworks are applied to real data from several AP exams with different levels of format effects.

The major findings based on the real data analyses are summarized as follows:

- The two MIRT models, SS-MIRT and BF-MIRT, tend to provide psychometric-property estimates that are similar to each other; and the estimates based on the

UIRT model are quite different from those based on the two MIRT models when the data are multidimensional.

- Estimates of reliability, classification consistency, and classification accuracy for the UIRT model are almost always smaller than those for the two MIRT models.
- Fitted observed-score distributions based on the UIRT and SS-MIRT models are very similar, which yet, are quite dissimilar to those based on the BF-MIRT model—however, the discrepancies become smaller when data are less multidimensional.
- Discrepancies in the psychometric-property estimates among the three IRT models decrease when the format effects are smaller or data are less multidimensional.
- The general patterns of conditional psychometric-property indices tend to be similar for all three IRT models.
- Conditional psychometric-property indices for scale scores show different patterns compared to those for composite raw scores.
- D-method and M-method tend to produce the overall estimates of psychometric properties that are highly comparable.
- Conditional psychometric-property indices based on the M-method are less vertically scattered than the D-method.

Since the study involves real data only, there is no absolute criterion to determine which model is most suitable for the data. However, the general tendency for the MIRT models to fit data with larger format effects better than the UIRT model, coupled with the fact that the UIRT model behaves differently from the two MIRT models seems to suggest that the MIRT models would be more appropriate for data with substantial format effects, and in such cases, reliability and classification consistency/accuracy estimates under the UIRT model may be underestimated.

Three estimation methods in relation to the latent-trait distributions are discussed, and two of them (i.e., D- and M-methods) were used and compared in the real-data examples. The P-method (not considered in this study) uses each person’s theta estimates, which requires selection of a latent-trait scoring method. The P-method would be most useful if the purpose is to compute and report psychometric-property estimates for each individual examinee in the sample. Both the D- and M-methods can be used for computing overall psychometric-property estimates such as reliability. However, when it comes to the conditional estimates, the M-method is generally preferred especially when used with the MIRT models because,

with the M-method, (a) it is less likely that unrealistic combinations of theta values are drawn, and (b) fitting a polynomial regression on the conditional SEMs is easier and more sensible.

The procedures presented in this paper for estimating psychometric properties concentrate on mixed-format tests. However, the procedures are general enough to be applied to situations in which items are grouped into any set of fixed categories, e.g., sub-content domains. If a test is developed according to a table of specifications and the score-scale of interest is the (weighted) composite score, a MIRT framework would be more sensible and defensible than the UIRT framework. However, choosing an IRT framework should be an informed decision. Most importantly, the test specifications should be reviewed carefully to understand how items are developed and clustered together. Some empirical dimensionality assessments and model-fit analyses would be useful. Also, the primary purpose of the IRT application should be clearly identified, because the effects of using different IRT frameworks may not be the same for different applications—for example, estimating reliability vs. test form equating.

Since the data examined in the present study had only two item types, more research is needed to illustrate the procedures for estimating psychometric properties when there are more than two item formats or sub-content categories. One practical issue to deal with in MIRT is that, as the number of distinct item types or sub-domains increases, the number of theta combinations that are to be evaluated increases exponentially. Another context that can be considered in future research is the case where raw scores on each section are converted to scale scores, and the score-scale of interest at the test level is the composite of the multiple scale scores over all sections. The general framework discussed in this paper is still applicable to the situation; however, the specific computational process and final outcome may be quite different.

References

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*, 3-14.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Brennan, R. L. (2011). *Utility indexes for decisions about subscores*. (CASMA Research Report No. 33). Iowa City, IA: CASMA, University of Iowa.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59*, 5-24.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, 29, 253-271.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612.

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

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, College Park, MD.

College Board. (2014). *English Language and Composition course description*. New York, NY: Author.

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.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*, 354-378.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D., Frank, E., Grochocinski, V., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204-229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*, 79-95.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note.

Hendrickson, A., Patterson, B., & Ewing, M. (2010). *Developing form assembly specifications for exams with multiple choice and constructed response items: Balancing reliability and validity concerns*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver.

Huynh, H. (1990). Computation and Statistical Inference for Decision Consistency Indexes Based on the Rasch Model. *Journal of Educational Statistics, 15*, 353-368.

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education, 10*, 359-375.

Kim, S. Y., & Lee, W. (2016). Classification consistency and accuracy for mixed-format tests. In M. J. Kolen, & W. Lee (Ed.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 4)* (CASMA Monograph No. 2.4) (pp. 113-147). Iowa City, IA: CASMA, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Knupp, T. L. (2009). *Estimating decision indices based on composite scores* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). American Council on Education & Praeger: Westport, CT.

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*, 15-24.

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.

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

LaFond, L. J. (2014). *Decision consistency and accuracy indices for the bifactor and testlet response theory models* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Lee, G., & Lee, W. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29*, 224-241.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1-17.

Lee, W. (2015). *MIRT-PP: Multidimensional item response theory for psychometric properties* [Computer software]. Iowa City, IA: CASMA, University of Iowa.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 33*, 129-140.

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: CASMA, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412-432.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

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

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

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Peterson, J. L., & Lee, W. (2014). Multidimensional item response theory observed score 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 3)*

(CASMA Monograph No. 2.3) (pp. 235-293). Iowa City, IA: CASMA, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

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.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163-184.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). American Council on Education & Praeger: Westport, CT.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. *Applied Psychological Measurement, 23,* 347-362.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30,* 29-40.

Steiger, J. H., & Lind, J. (1980). *Statistically-based tests for the number of common factors*. Paper presented to the Annual Meeting of the Psychometric Society, Iowa City.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29-44). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37,* 141-162.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54,* 427-450.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46,* 177-197.

Zhang, M., Kolen, M. J., & Lee, W. (2014). A comparison of test dimensionality assessment approaches 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) (pp. 161-200). Iowa City, IA: CASMA, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Table 1

Descriptive Statistics for Three Exams

Statistics	Spanish	English	US History
Composite Raw Score Range	0-149	0-135	0-161
# of MC Items	65	54	80
# of FR Items (maximum score)	6 (9,5,9,5,9,5)	3 (9,9,9)	3 (9,9,9)
N	6,000	6,000	6,000
Mean	63.186	49.025	55.213
SD	14.216	12.448	19.454
Skewness	-.494	-.470	-.169
Kurtosis	.320	-.225	-.868
w_{MC}: w_{FR}	1:2	1:3	1:3
$\hat{\rho}_{\theta_{MC}\theta_{FR}}$	.758	.799	.917
$\hat{\rho}_{T_{MC}T_{FR}}$	.760	.811	.912
Cut Scores	74/84/98/109	52/73/88/100	45/72/86/107

Note. *N* = sample size; SD = standard deviation; w_{MC} = weight for MC section; w_{FR} = weight for FR section; $\hat{\rho}_{\theta_{MC}\theta_{FR}}$ = estimated IRT latent-trait correlation; $\hat{\rho}_{T_{MC}T_{FR}}$ = estimated classical disattenuated correlation.

Table 2

Fit Statistics

Model	AIC	BIC	RMSEA	$S-X^2$ ($\alpha = .05$) # of Misfit Items
Spanish				
UIRT	548861.43	550489.42	0.11	9
SS-MIRT	546447.49	548082.17	0.11	8
BF-MIRT	542050.18	544153.82	0.13	7
English				
UIRT	408460.61	409746.92	0.09	14
SS-MIRT	404053.24	405721.41	0.10	15
BF-MIRT	407724.74	409017.74	0.09	13
US History				
UIRT	594730.97	596539.84	0.12	3
SS-MIRT	594236.64	596052.21	0.12	3
BF-MIRT	587857.53	590222.46	0.12	10

Table 3

Estimated Overall SEM and Reliability for Composite Raw Scores

Exam	Overall SEM			Reliability		
	UIRT	SS-MIRT	BF-MIRT	UIRT	SS-MIRT	BF-MIRT
D-method						
Spanish	7.278	6.771	6.226	.872	.890	.908
English	8.083	7.352	7.327	.823	.860	.857
US History	7.298	6.868	6.794	.927	.936	.936
M-method						
Spanish	7.263	6.766	6.219	.873	.892	.909
English	8.081	7.348	7.327	.825	.860	.854
US History	7.293	6.863	6.794	.927	.935	.935

Table 4

Estimated Overall SEM and Reliability for Scale Scores

Exam	Overall SEM			Reliability		
	UIRT	SS-MIRT	BF-MIRT	UIRT	SS-MIRT	BF-MIRT
D-method						
Spanish	3.596	3.342	3.084	.872	.890	.909
English	4.200	3.835	3.816	.817	.855	.852
US History	3.250	3.065	3.024	.923	.933	.933
M-method						
Spanish	3.604	3.343	3.091	.873	.892	.910
English	4.200	3.831	3.816	.818	.854	.849
US History	3.246	3.057	3.024	.923	.932	.933

Table 5

Estimated Overall Classification Consistency and Accuracy

Index	D-method			M-method		
	UIRT	SS-MIRT	BF-MIRT	UIRT	SS-MIRT	BF-MIRT
Consistency $\hat{\phi}$						
Spanish	.550	.574	.605	.549	.578	.604
English	.519	.557	.558	.520	.558	.555
US History	.657	.676	.678	.658	.673	.678
Accuracy $\hat{\gamma}$						
Spanish	.653	.671	.699	.646	.675	.699
English	.633	.666	.667	.633	.668	.665
US History	.745	.764	.766	.750	.762	.766
False Positive Error Rate $\hat{\gamma}^{+}$						
Spanish	.182	.185	.172	.201	.185	.170
English	.170	.186	.186	.204	.185	.188
US History	.131	.128	.128	.138	.131	.129
False Negative Error Rate $\hat{\gamma}^{-}$						
Spanish	.165	.145	.129	.153	.141	.131
English	.197	.148	.147	.163	.147	.147
US History	.124	.108	.107	.112	.106	.105

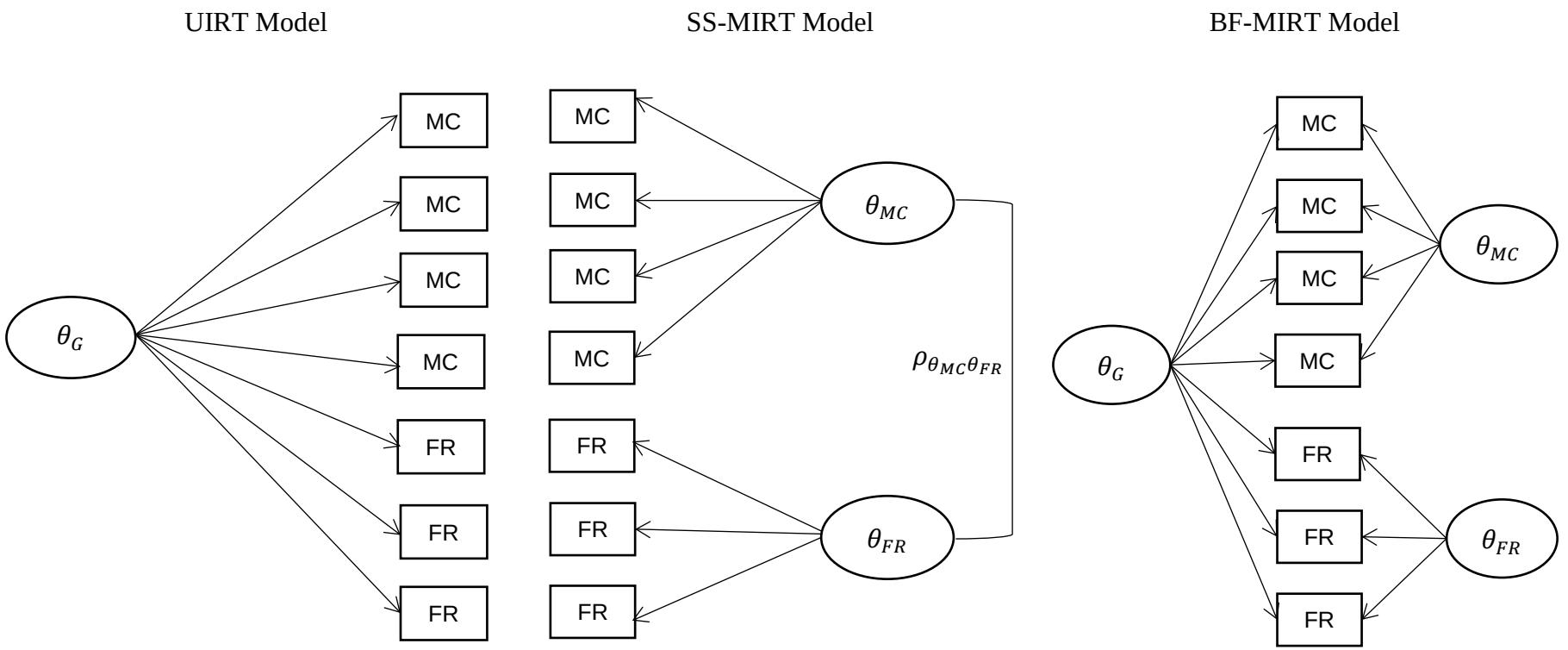


Figure 1. Diagrams for UIRT, SS-MIRT, and BF-MIRT.

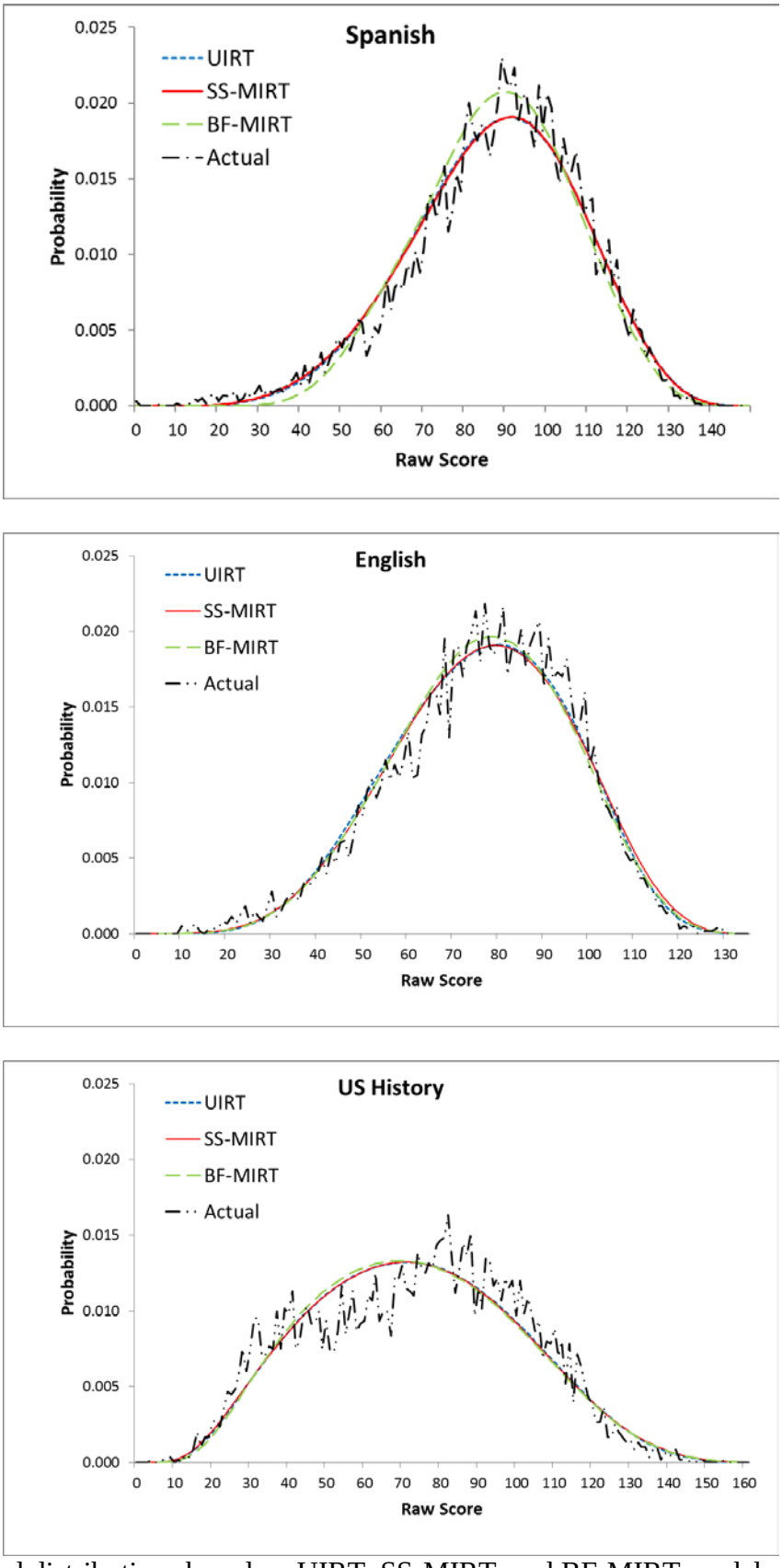


Figure 2. Fitted distributions based on UIRT, SS-MIRT, and BF-MIRT models.

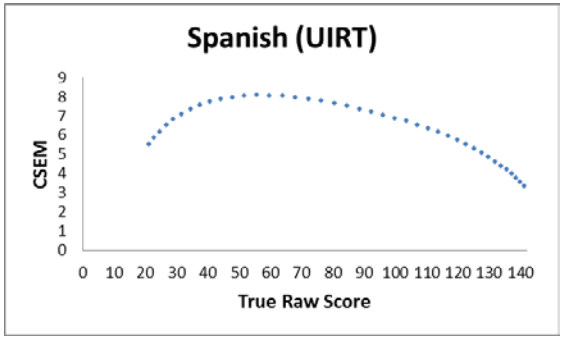
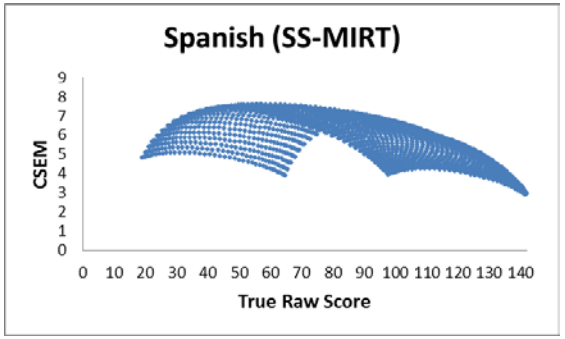
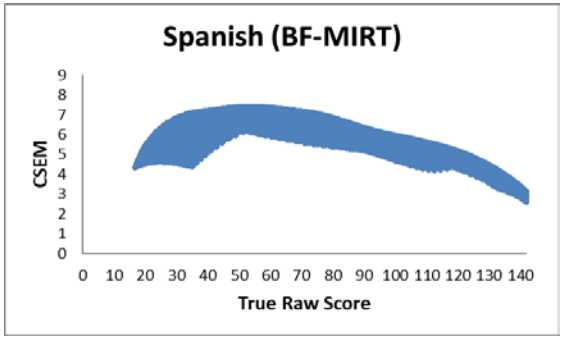
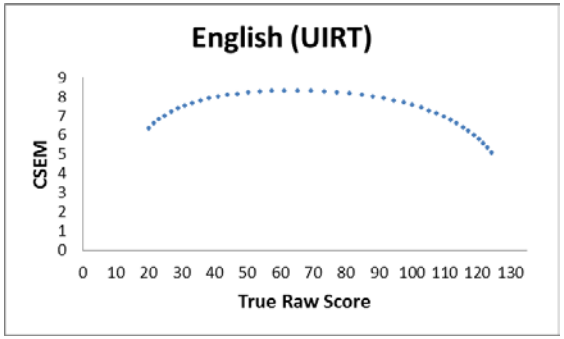
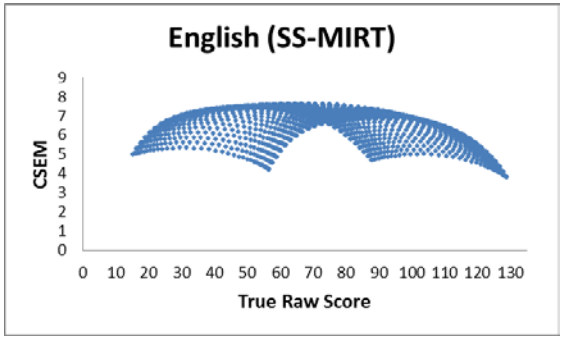
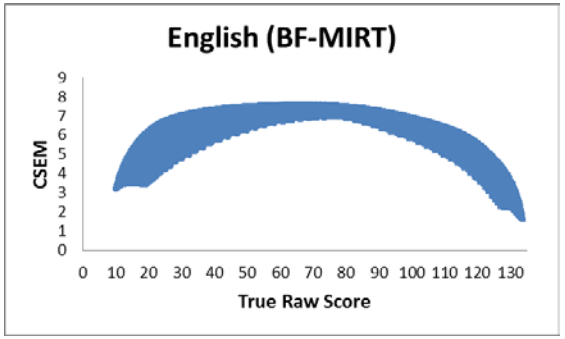
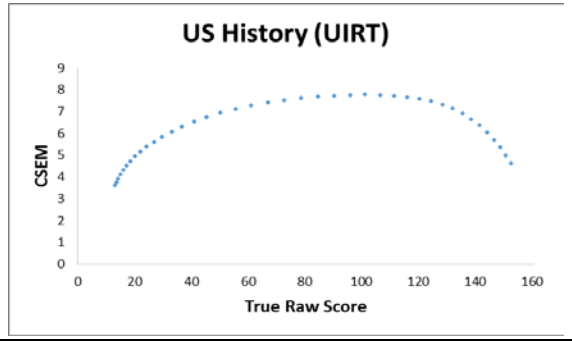
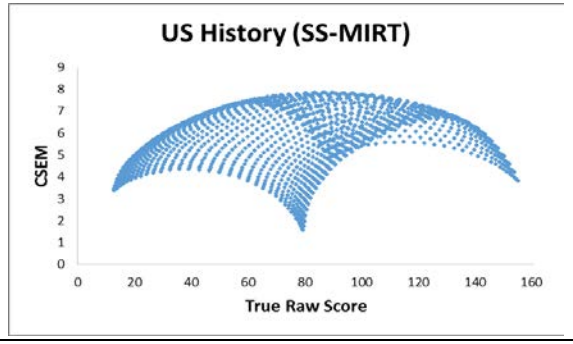
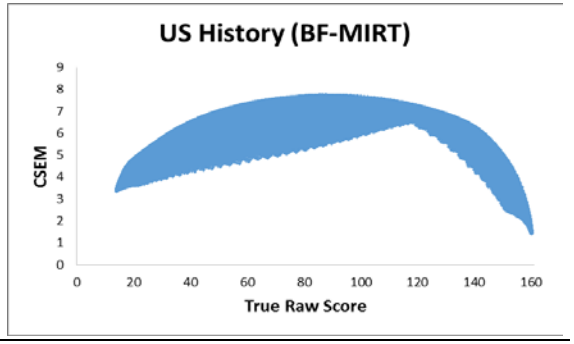
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 3. CSEMs for composite raw scores based on D-method.

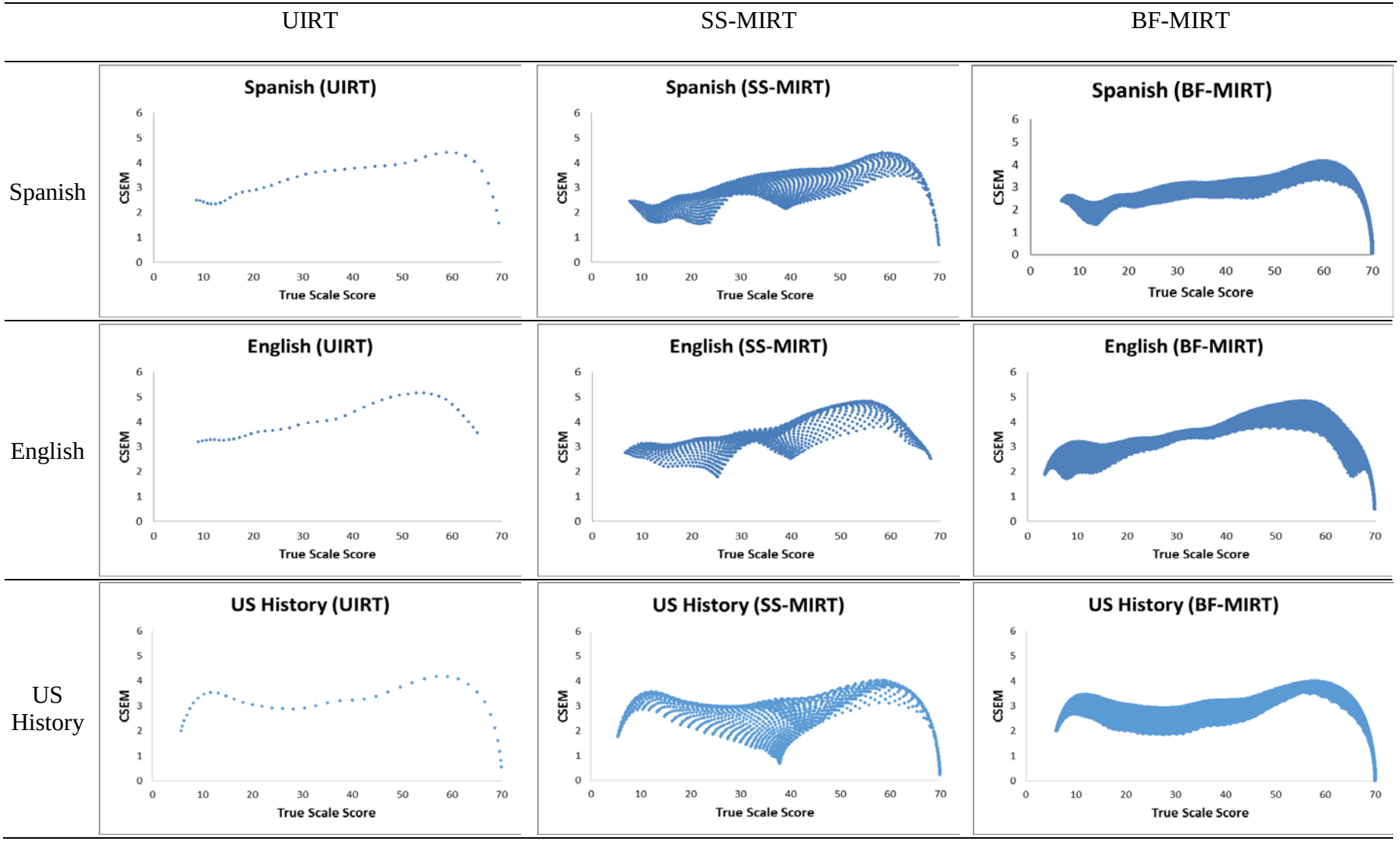


Figure 4. CSEMs for scale scores based on D-method.

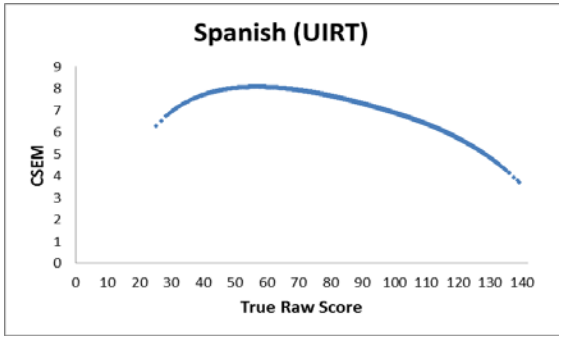
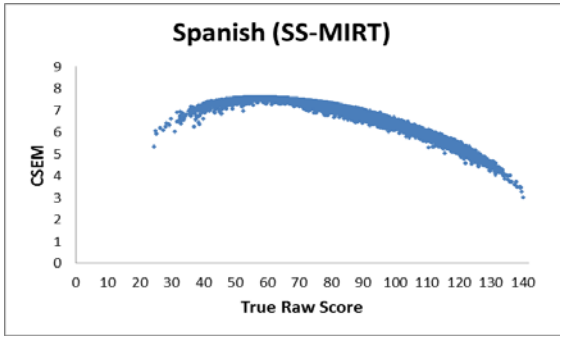
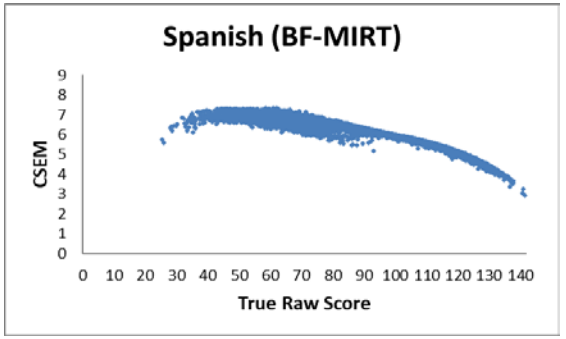
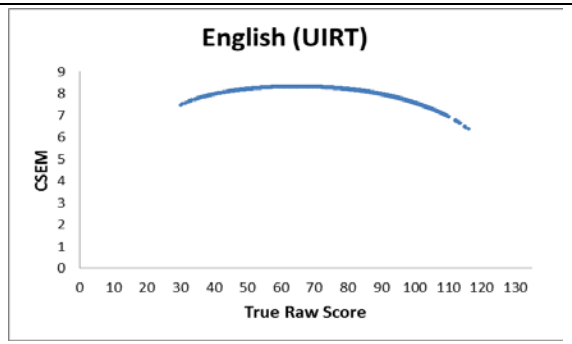
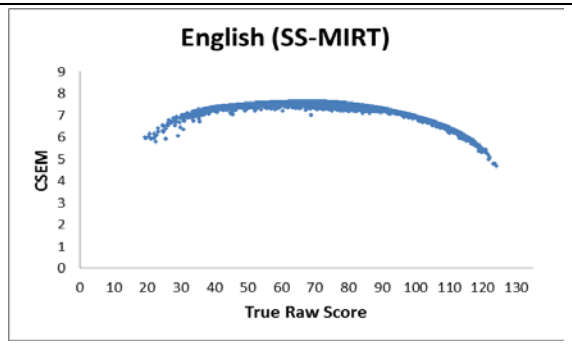
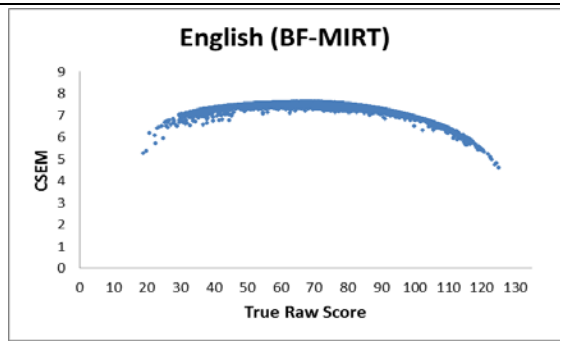
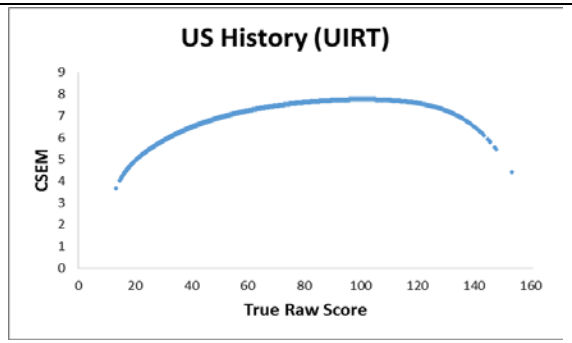
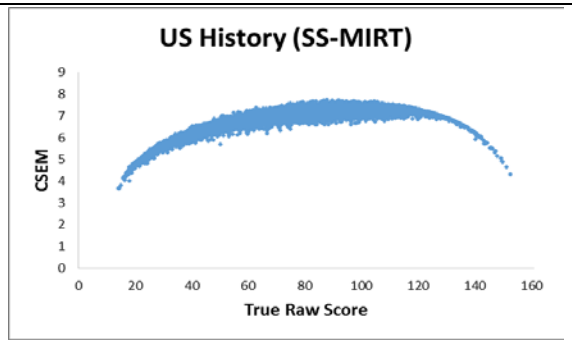
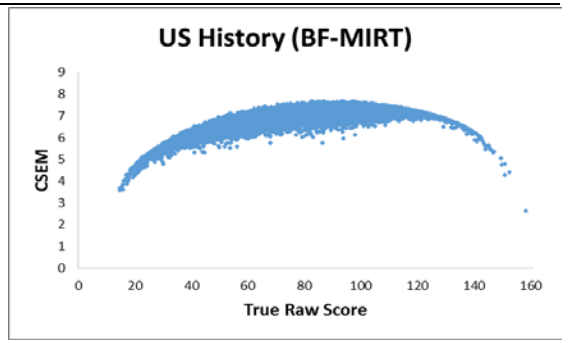
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 5. CSEMs for composite raw scores based on M-method.

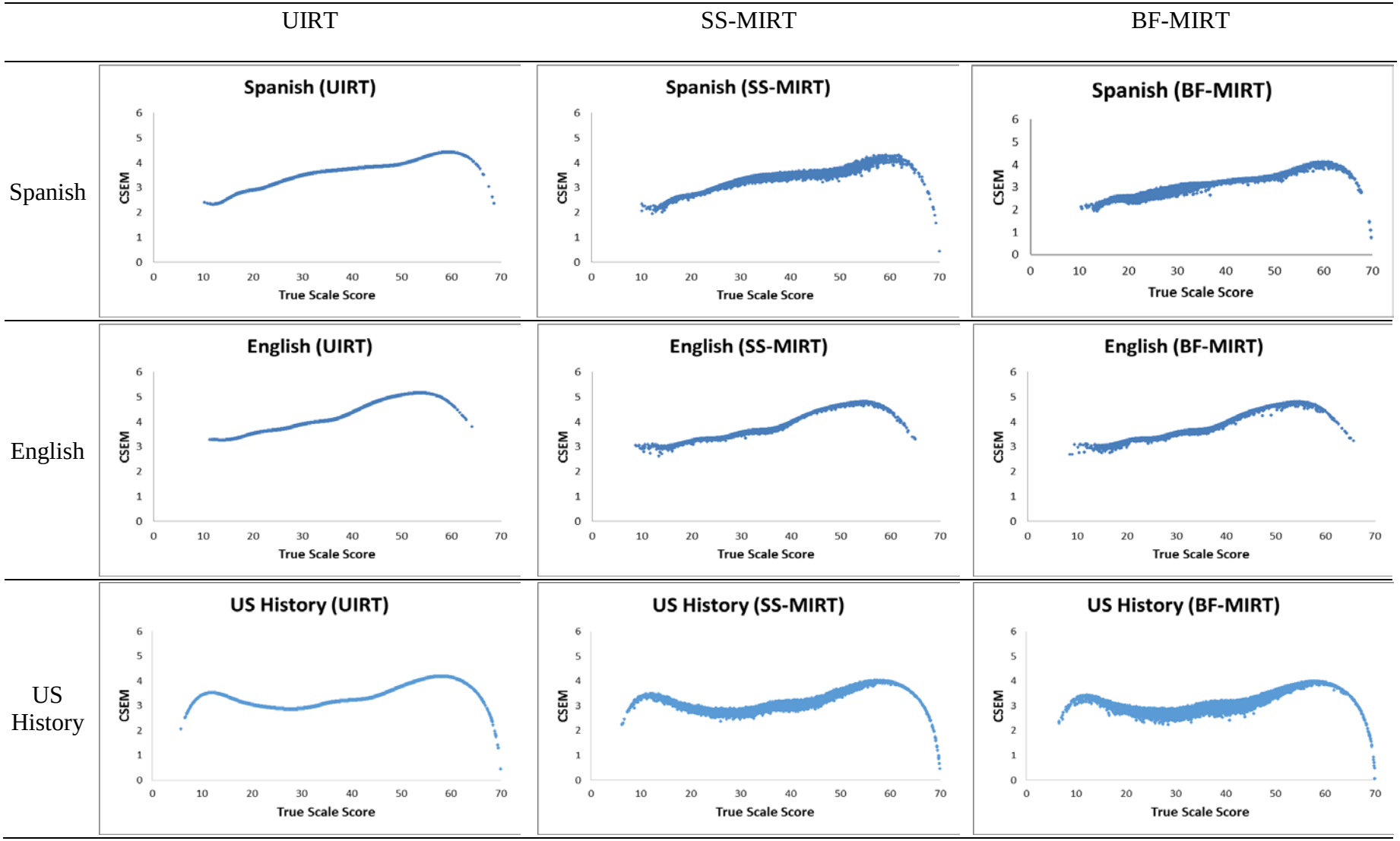


Figure 6. CSEMs for scale scores based on M-method.

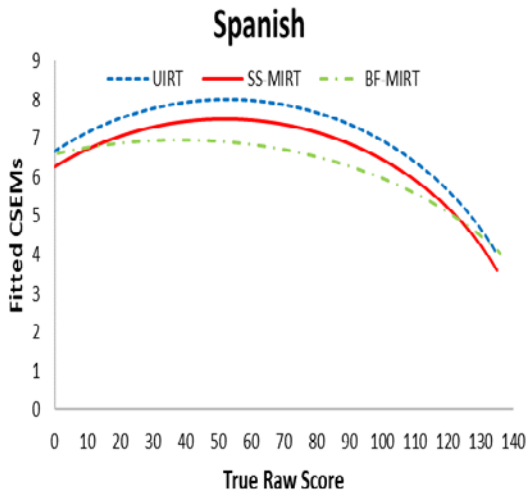
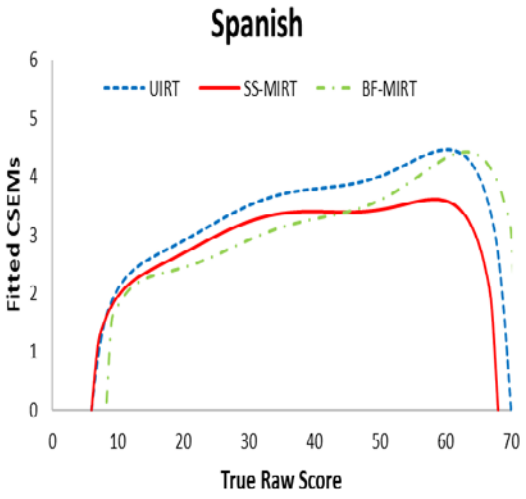
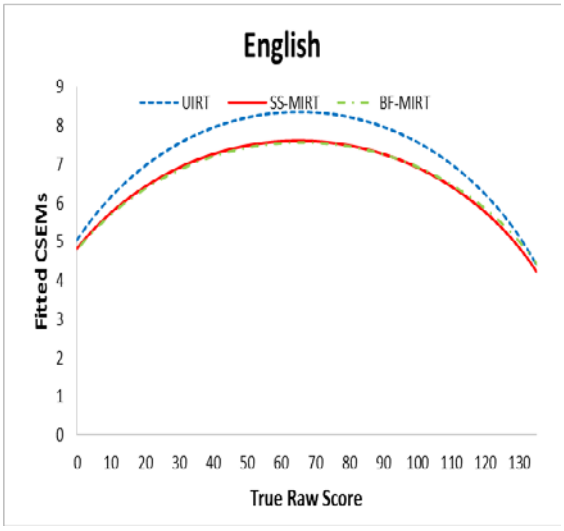
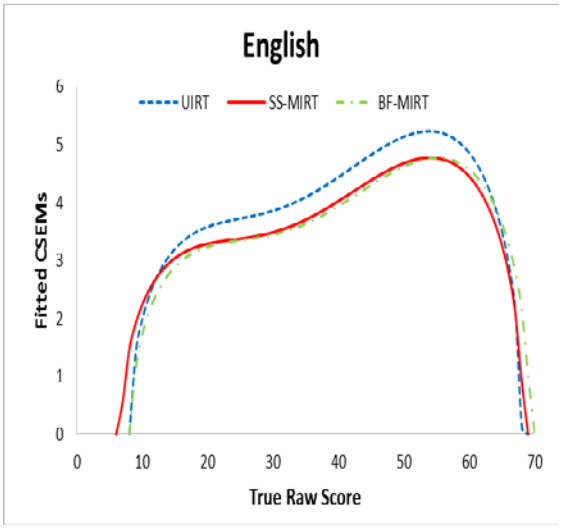
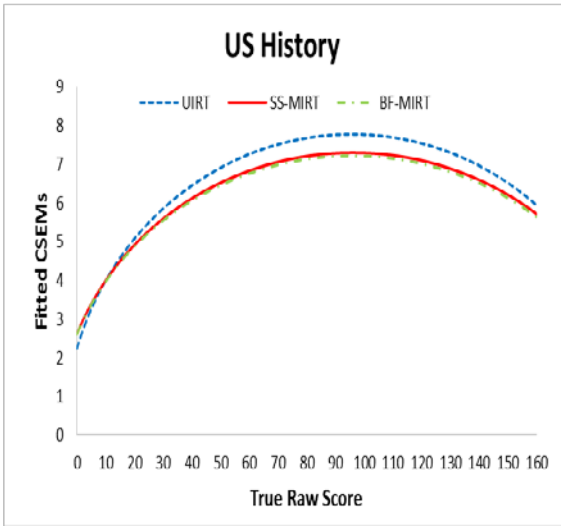
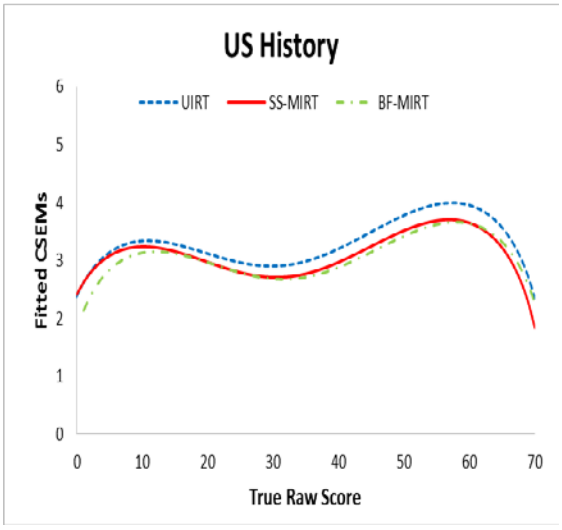
	Composite Raw Scores	Scale Scores
Spanish		
English		
US History		

Figure 7. Fitted CSEMs for composite raw and scale scores based on M-method using a polynomial regression.

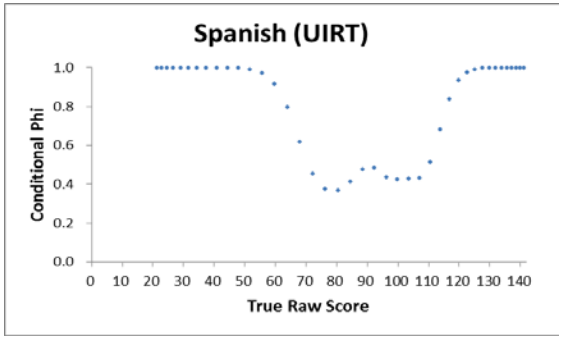
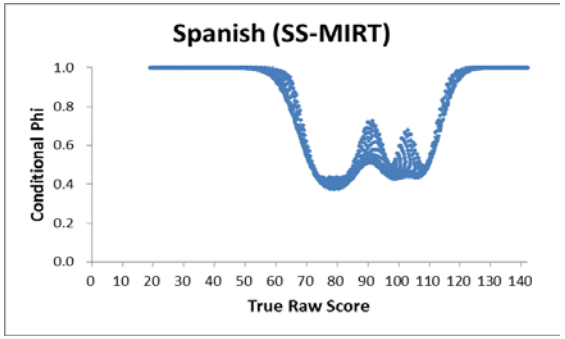
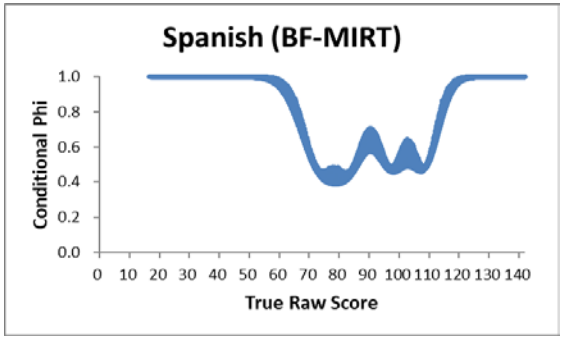
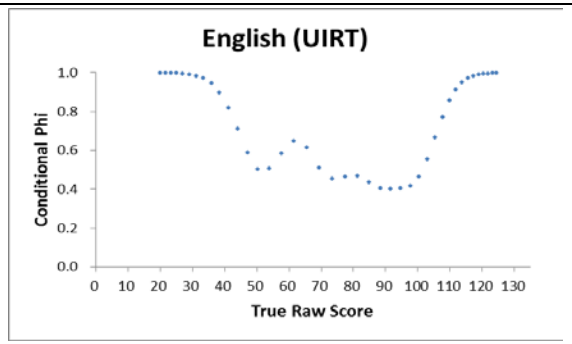
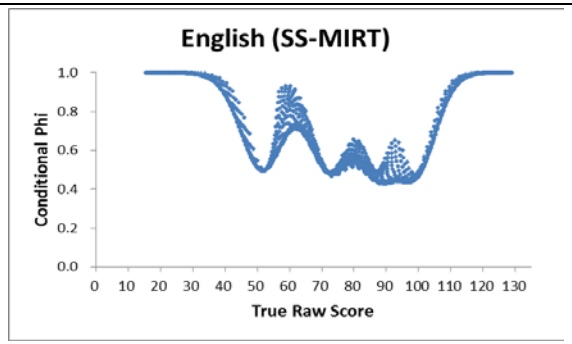
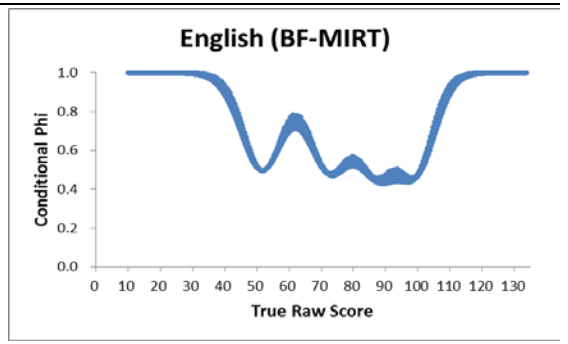
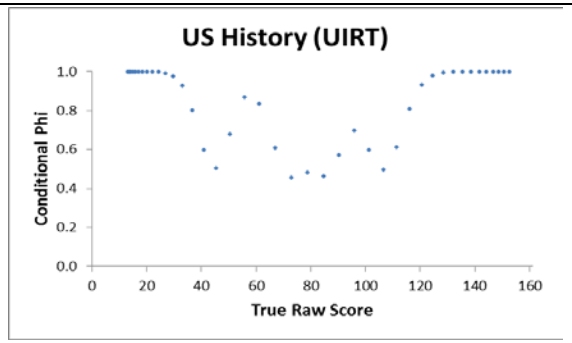
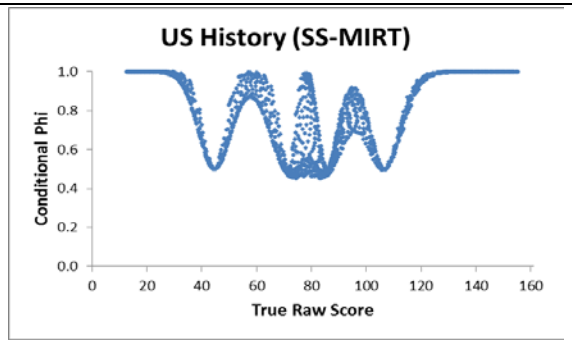
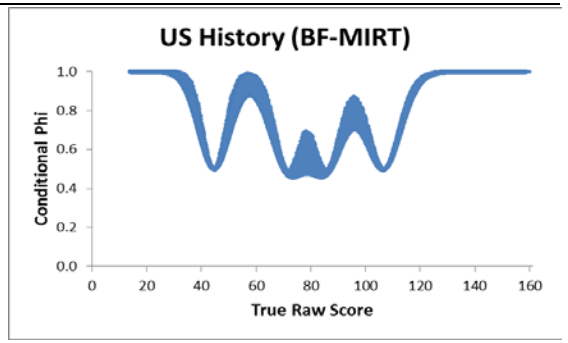
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 8. Conditional classification consistency based on D-method.

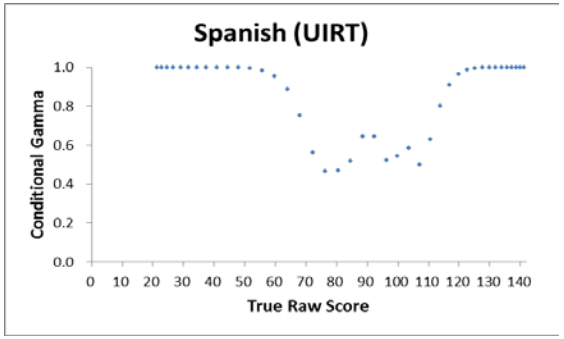
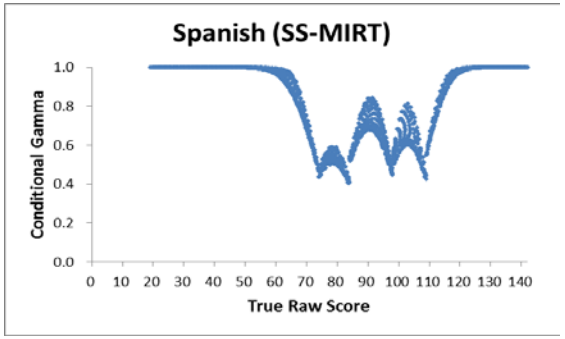
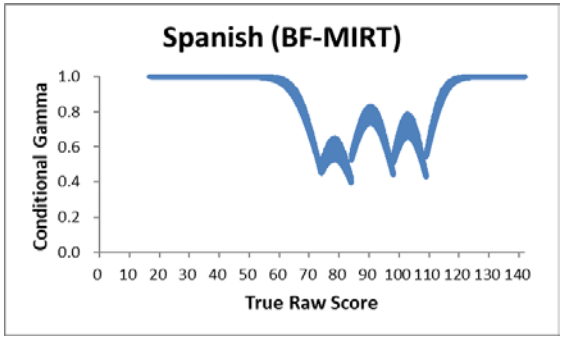
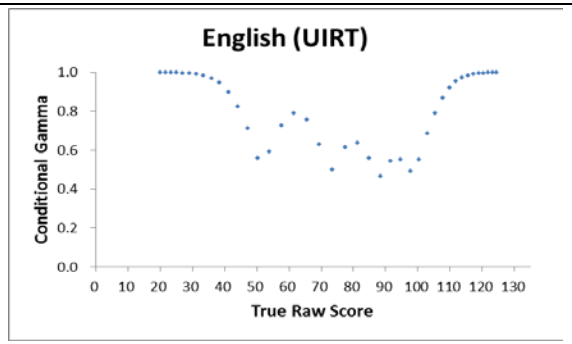
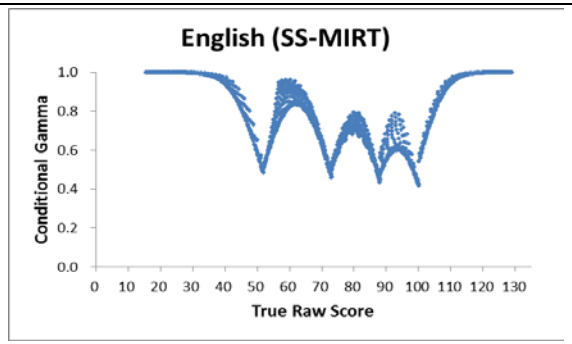
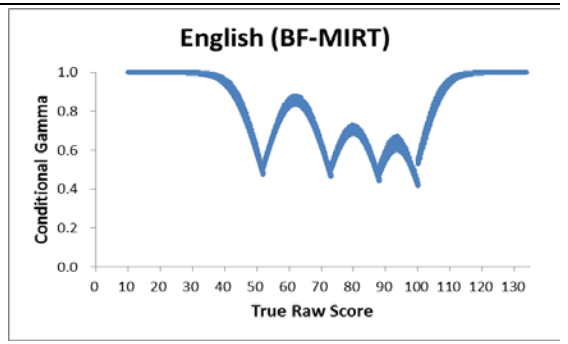
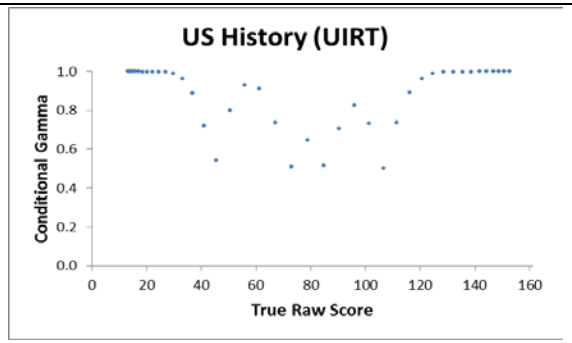
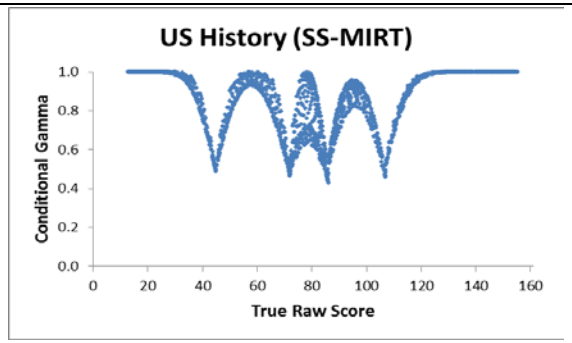
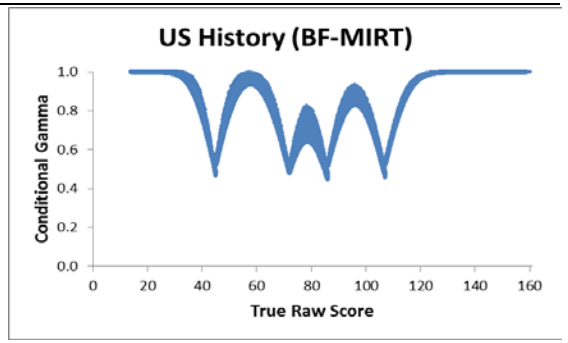
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 9. Conditional classification accuracy based on D-method.

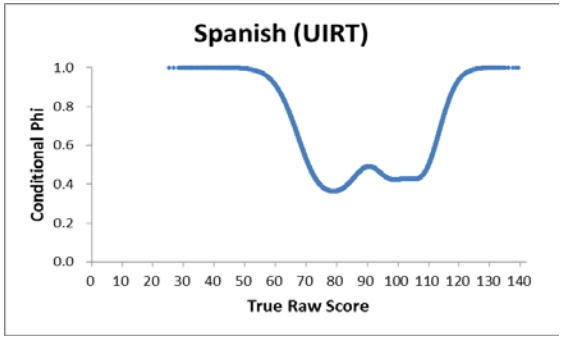
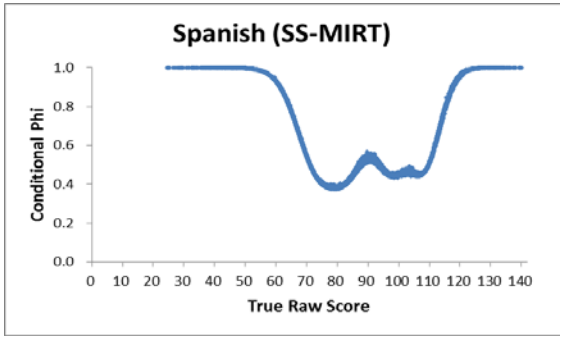
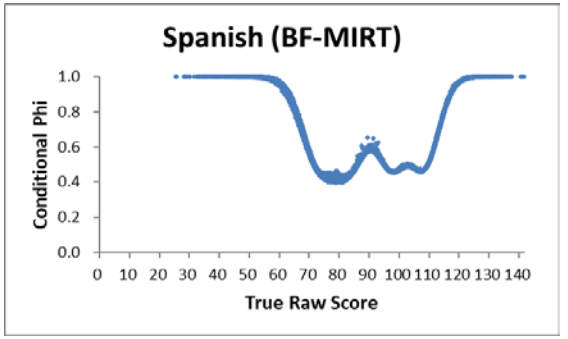
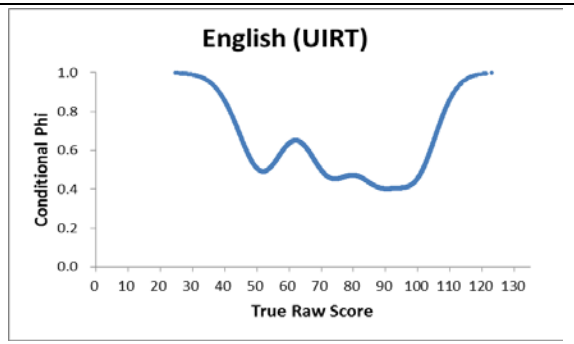
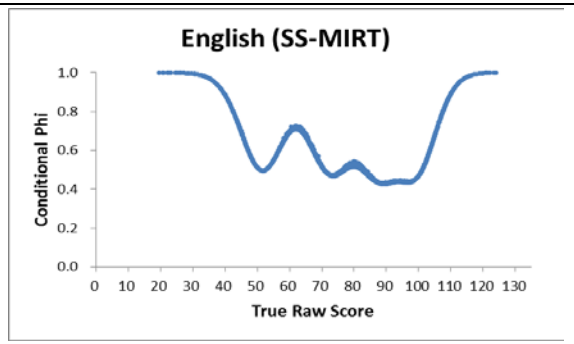
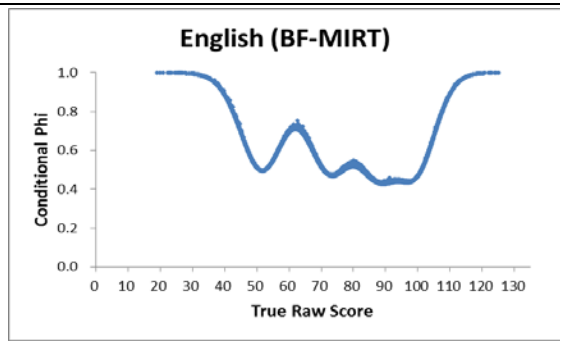
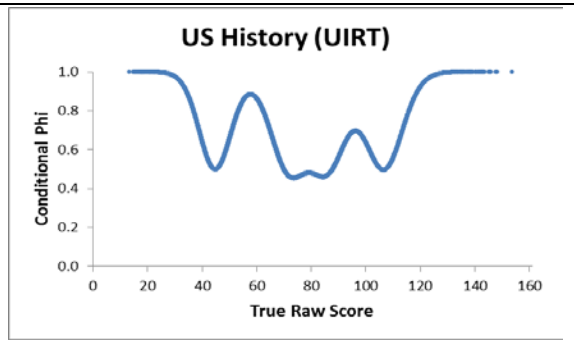
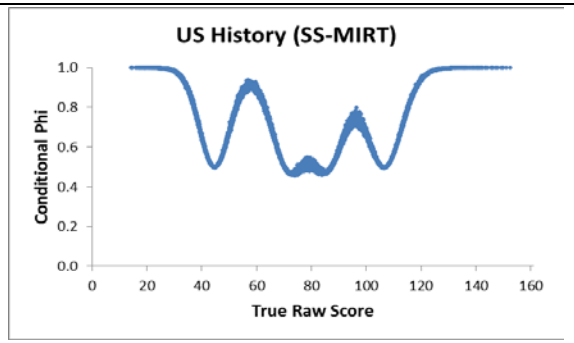
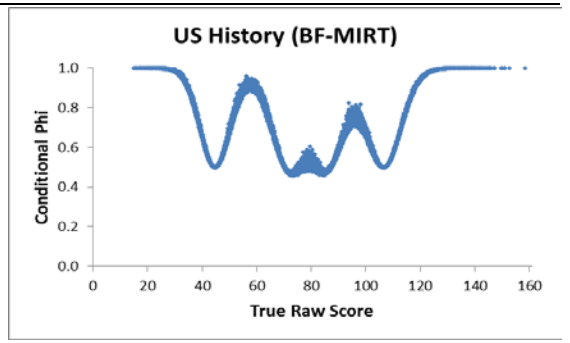
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 10. Conditional classification consistency based on M-method.

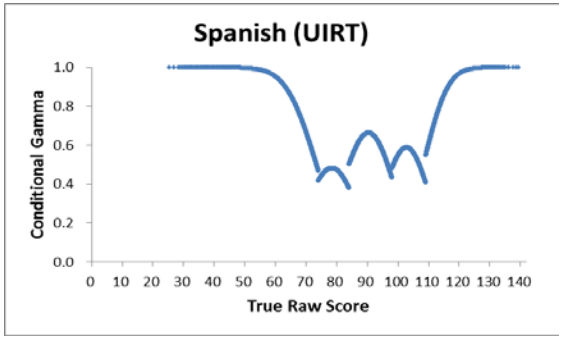
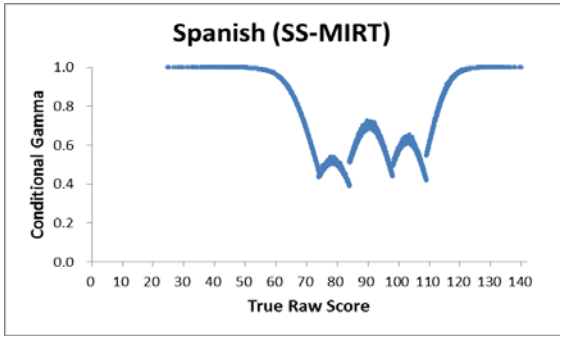
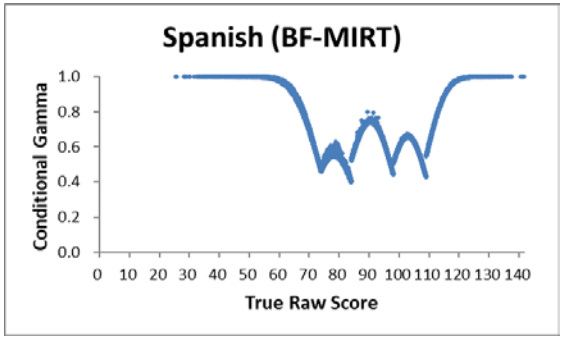
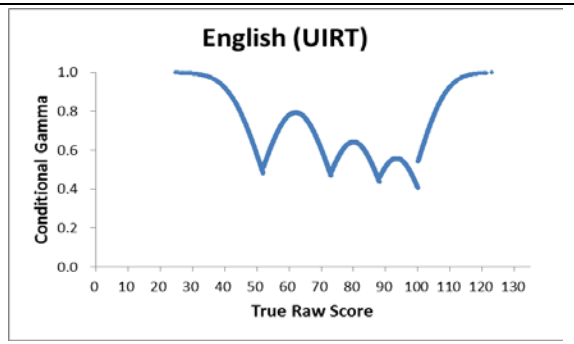
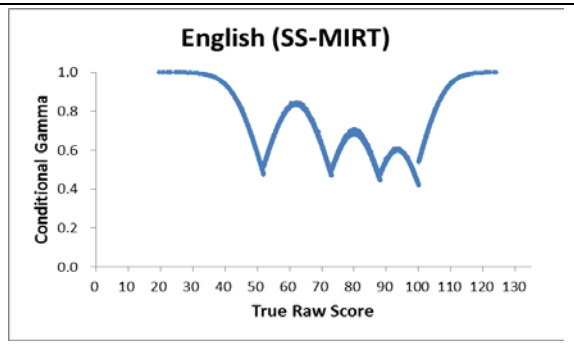
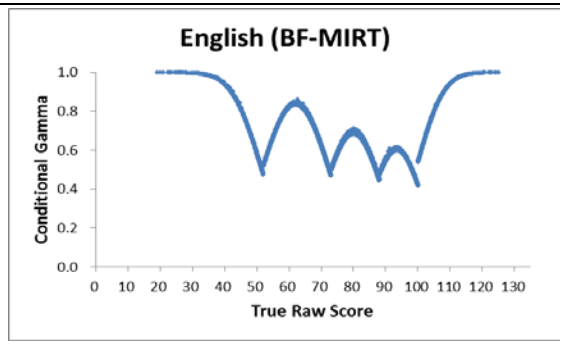
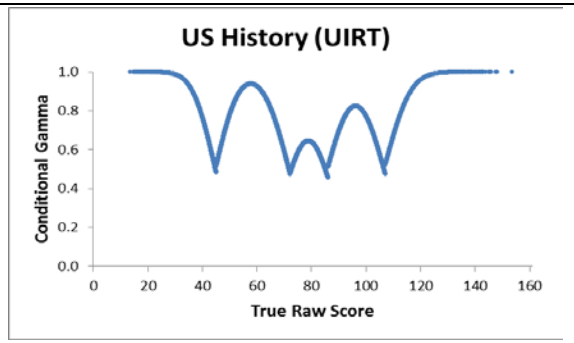
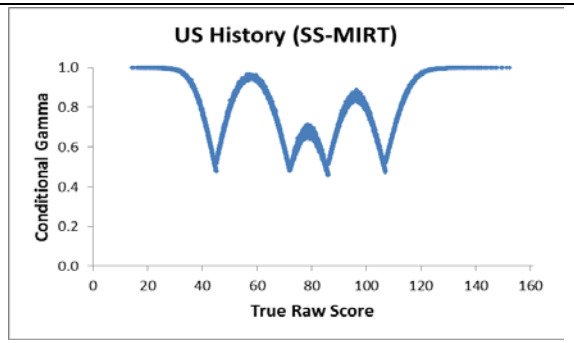
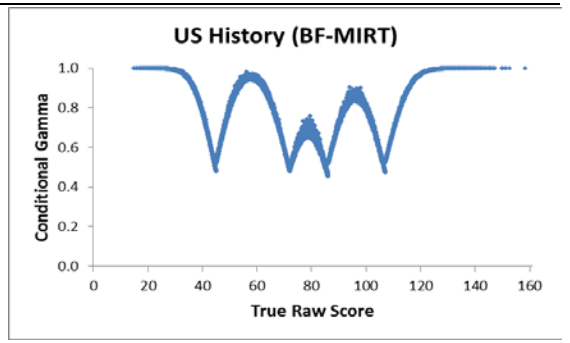
	UIRT	SS-MIRT	BF-MIRT
Spanish			
English			
US History			

Figure 11. Conditional classification accuracy based on M-method.

Chapter 3: Can Task Models be Used for Equating Purposes?

Jaime L. Malatesta and Huan Liu

The University of Iowa, Iowa City, IA

Abstract

The main purpose of this research was to investigate whether free-response (FR) items generated from the same task model could be used to improve equating. More specifically, FR items built from the same task model were evaluated as if they were traditional common items and based on the results, several combinations of FR item(s) were added to the operational anchor set and used for scale linking. Equating was then conducted under common item nonequivalent groups design. Pseudo-groups data from three Advanced Placement world language tests (German, French, and Italian) than span four administration years were used. Studied conditions included type of equating method, level of group ability differences, and anchor composition. In general, findings from this study generally do not support the use task model-derived FR items as anchor items for equating until further research is conducted.

Can Task Models be Used for Equating Purposes?

Mixed-format tests that contain both multiple-choice (MC) and free-response (FR) items have become increasingly popular in the educational testing field. MC and FR items have different characteristics; the advantages of MC items are that they can assess broad content coverage and be efficiently scored whereas the advantage of FR items is they can typically measure higher order cognitive skills (Hagge, 2010). However, the combination of two item formats in a single test also makes it more complicated to conduct certain psychometric procedures, such as equating.

According to Kolen and Brennan (2014), in the common-item nonequivalent group (CINEG) equating design, common-items (also referred to as anchor items in this study) need to represent the content and statistical characteristics of the total test. However, because FR items are easy to memorize, including them as part of the anchor set can consequently affect equating accuracy. Therefore, the decision of whether or not to include FR items in the anchor set often represents a tradeoff between adhering to strict content representativeness and proactively guarding against inflated equating error associated with exposed FR anchor items.

The use of task models could provide one solution to this problem. A task model, in simplest words, is a collection of relevant tasks or item features, where each feature has a defined set of possible variations. More specifically, each task model integrates declarative knowledge components, relationships among those components, and cognitive skills, as well as relevant content, contexts, and auxiliary features that affect the cognitive complexity of the task (Luecht, 2013). Furthermore, task models must have a relative or absolute location on the ability scale. Task models that demand a higher level of reasoning and understanding from the examinee would be located higher on the ability scale relative to those that require a lesser amount. Last, task models should contain specifications and features that could be used to easily and quickly generate a family of related items or items that measure the same specific construct.

Task models can be beneficial in several testing circumstances. First, task models allow for the efficient construction of a large number of items that meet specific requirements. Second, items built from the same task model usually share similar content as well as cognitive and psychometric properties. In an ideal situation, task models could generate a collection of items with such similar statistical characteristics that they could be used as anchor items in linking and equating procedures.

Malatesta and Liu (2016) adapted several commonly used methods that were originally designed to detect unstable MC anchor items as well as one method used to detect differential item functioning, to evaluate the interchangeability of FR items generated from the same task model. Specifically, four approaches were used, including: squared differences of item characteristic curves, the Robust z method (Huynh, 2000), ordinal logistic regression, and visual inspection of IRT item parameter estimates. Their rationale for applying these criteria was that task model derived items are generally developed with the intent of being mutually interchangeable with respect to content and psychometric properties (Bejar, 2002).

Using the aforementioned criteria, Malatesta and Liu (2016) identified the best and worst performing FR item for each exam. The label “best” was reserved for the FR item that generally performed most similarly across forms whereas the “worst” item performed most differently. These labels are carried into the current study.

In their study, Malatesta and Liu (2016) suggested that the best FR items could possibly be treated as anchor items for equating purposes since they behaved similarly to the actual MC anchor items in their study. By including FR items in the anchor set, the content representativeness would be improved which might improve equating accuracy. The current study is a continuation of Malatesta and Liu’s (2016) study, and compares the equating relationships resulting from including the best FR item, worst FR item, and all FR items in the anchor set. Results from these aforementioned conditions are compared with those found by using the operational MC-only anchor set.

In summary, the existing literature and previous studies show that items generated from the same task model have the potential to be treated as anchor items. This study mainly aims to evaluate whether the FR items built from the same task model should be included in the anchor set under the CINEG design. More specifically, this research intends to gather evidence concerning whether task model generated FR items should be treated as anchor items and compares the equating results using different anchor sets: only MC anchor items, MC plus the best FR item, MC plus the worst FR item, and MC plus all FR items. Additional factors such as group ability differences are also taken into consideration.

The remaining report is organized into three major sections. The method section describes the procedure used to create the pseudo-group data as well as the equating methods used and subsequent evaluation criteria. In the results section, the overall findings using all

evaluation criteria are presented across three AP exams, separately. Last, general conclusions, limitations, and future research ideas are presented in the discussion section.

Method

Data

Three Advanced Placement (AP) World Language and Culture exams were used in this study: French, German, and Italian. Each exam is mixed-format and contains either 65 or 70 MC items, along with 4 FR items. The MC items are scored dichotomously and the FR items are each scored on a 6-point scale (i.e., 0, 1, 2, 3, 4, and 5). The first FR item is an interpersonal writing task, the second is a presentational writing task, the third is a culmination of five short interpersonal speaking tasks aggregated into one score, and the fourth FR is a presentational speaking task.

Four years of AP data were included in this study: 2012, 2013, 2014, and 2015. Due to the linking design, only main forms were included in the study which resulted in three links per subject: 2013 linked to 2012, 2014 linked to 2012, and 2015 linked to 2013. Using the operational samples, pseudo-groups were created in order to achieve various levels of group ability differences (defined by examinee total scores on the operational anchor set). Group differences are discussed in more detail later.

The descriptive statistics of unweighted composite raw scores for the pseudo groups' data that were used in this study are presented in Tables 1- 9. It is important to note that even though the fourth FR item was found to perform the worst across all three exams, even the best performing FR item showed different statistical characteristics across the three exams (Malatesta & Liu, 2016). Furthermore, the disattenuated correlations between MC and FR section scores varied slightly across exams (see Table 10). Disattenuated correlations are loosely used to evaluate the extent to which MC and FR items measure the same construct.

Factors of Investigation

Three factors of investigation were considered: composition of anchor set used for scale linking, degree of group ability differences, and equating method.

Anchor composition. To investigate the performance of the best FR items, four anchor compositions were considered in this study: MC items only, MC items plus the best FR item, MC items plus the worst FR item, and MC items plus all FR items. It should be noted that "MC

items" in the previous sentence refers to the MC anchors that were used to conduct the operational equating.

Based on the study by Malatesta and Liu (2016), each of the four FR item types performed differently depending on the subject area, except for FR 4 (presentational speaking) which behaved most differently across forms for all exams and therefore was labeled as the worst FR item. For German, FR 1 – FR 3 all performed relatively well in contrast to FR 4; however FR 2 performed the best. For Italian, FR 1, FR 2, and FR 4 all performed relatively poorly compared with FR 3, which was selected as the best performing FR item. For French, FR 1 and FR 2 performed similarly, but strictly speaking, FR 1 performed the best.

Group ability difference. In this study, ability differences between the old and new form groups were created using a sampling process based on the demographic variable parental education level, which represented the highest level of education that either the examinee's mother or father had achieved (coded 0-6). Group ability difference was defined as the effect size (*ES*) of the difference in total scores (calculated as a simple sum of MC anchor items) between the old and new groups. A positive relationship was observed between parental education levels and total scores on the anchor items. Therefore, different effect sizes could be obtained across old and new forms by oversampling certain demographic groups using an iterative procedure. Under the CINEG design, three levels of group differences were considered: 0, 0.1, and 0.3. *ES* was calculated using the following equation:

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2}}}, \quad (1)$$

where $\bar{x}$ represents the mean anchor total score; n is the number of examinees; s^2 is the variance of anchor total scores; and subscripts 1 and 2 correspond to the new and old form groups, respectively.

Linking and equating methods. Equating was conducted in accordance to the CINEG design using four methods: frequency estimation (FE), chained equipercentile equating (CE), IRT true-score equating (TS), and IRT observed-score equating (OS). All equating methods were conducted using *Equating Recipes* (Brennan et al., 2009). A conversion table was created for the old form, in which raw scores were converted to unrounded normalized scale scores that ranged from 0 to 70 and to AP Grades of 1-5. The normalized score scale was constructed using the

method described by Kolen and Brennan (2014) and had a mean of 35 and standard deviation of 10.

For the two traditional equating methods (i.e., FE and CE), bivariate log-linear presmoothing was conducted using 6 degrees of smoothing for the total scores and CI scores with one cross-product (also referred to as the 6, 6, 1 model). The smoothed distributions appeared to fit the observed data well. Equating was always conducted such that the synthetic population was defined as the new-form group.

For the IRT equating methods, item and examinee ability parameter estimates were obtained using flexMIRT 3.0 (Cai, 2015). The three-parameter logistic model (3PL; Birnbaum, 1968) was used for the MC items, and Muraki's (1992) generalized partial credit (GPC) model was used for FR items. The flexMIRT default settings were used with two exceptions: (1) results for the 3PL model were requested in the normal metric using the keyword "Normalmetric3PL = yes" and (2) a prior distribution (i.e., c ~ beta (1, 4)) was used to estimate the pseudo-guessing parameter.

Using the item parameter estimates from flexMIRT, the computer program STUIRT (Kim & Kolen, 2004) was used to estimate the Haebara (1980) linking coefficients. Research has shown that the Haebara and Stocking-Lord (1983) methods are equally accurate (Hanson & Beguin, 2002; Kim & Kolen, 2004a; Kim & Lee, 2006) and therefore the decision to use the Haebara method was arbitrary. Different coefficients were obtained depending on which anchor composition was used. The linking coefficients were then applied to the new form item parameter estimates in order to place them on the old form scale. After scale transformation, equating was conducted using *Equating Recipes* (Brennan et al., 2009).

Evaluation Criteria

Since this study used real data, the true equating relationship was not known which resulted in the need to create a reasonable criterion. In this study, data were transformed to approximate the random group design using pseudo-group methodology. Under the RG design, parental education level was again used to randomly sample old and new form groups that had effect size differences of less than 0.01 with respect to anchor total scores. An attempt was made to create pseudo-groups with large sample sizes; however due to the nature of the operational data, this was not always achievable. Descriptive statistics for the RG pseudo-group samples can be found in Table 11.

Each studied equating condition was evaluated against a criterion relationship that employed the same equating method. For example, RG traditional equipercentile equating with log-linear presmoothing served as the criterion equating relationship for the studied conditions involving the FE and CE methods. Whereas, RG IRT OS or IRT TS served as the criterion relationship for studied conditions that involved the IRT OS or IRT TS methods, respectively. Even though the criterion equating relationships varied depending on the studied method, the same RG pseudo-group samples were used to obtain all criterion equating relationships.

Several indices were used to evaluate the equating relationships stemming from use of different anchor compositions, group ability differences, and equating methods, and will be discussed next.

wRMSD. First, weighted Root Mean Squared Difference (wRMSD) were computed to evaluate the discrepancies between the studied equivalents and the criterion equivalent for unrounded raw composite scores and unrounded scale scores. The wRMSD can be expressed as

$$\text{wRMSD} = \{\sum_{j=0}^{z} w_j [eq_Y(x_j) - eq_{Y_C}(x_j)]^2\}^{1/2} . \quad (2)$$

Here, $eq_Y(x_j)$ represents the equated equivalent for raw score *j* using one of the studied equating conditions, $eq_{Y_C}(x_j)$ represents the criterion equivalent of raw score *j*, w_j is the relative frequency associated with raw score *j* on the new form, and *Z* is the maximum raw composite score for the new form. Equation 2 was also applied to unrounded scale scores by replacing raw scores with scale scores. Smaller values indicate a higher degree of agreement between the studied conditions and the criterion.

Difference That Matters. The differences between the studied equivalents and criterion equivalents were evaluated using the Difference That Matters (DTM; Dorans, Holland, Thayer, & Tateneni, 2003) criteria. According to Dorans et al. (2003), a DTM is marked by a difference of half a reported score unit. The rationale of the DTM is that if two unrounded scale scores are within half a unit of each other, they may ultimately receive the same reported score after rounding. Thus, a difference of less than half a score unit might be considered acceptable. In the current study, discrepancies between the studied equating results and the criterion equating results were evaluated using the 0.5 benchmark for unrounded raw and scale scores. For each studied condition, the proportion of old form equivalents that fell within the DTM boundaries is reported.

Classification consistency for AP grades. Examinees who take the AP exams are classified into AP grades, ranging from 1 to 5. Most colleges require students to receive a score 3 or higher in order to receive college course credit(s). Therefore, classification consistency (CC) for AP grades has a lot of practical significance for stakeholders.

In this study, CC was calculated as the agreement between AP grades that the new form examinees received based on the criterion equating relationship compared with the studied condition. More specifically, CC for AP grades was computed as,

$$\mathrm{CC} = \frac{n_{11}+n_{22}+n_{33}+n_{44}+n_{55}}{N}. \quad (3)$$

Here, n_{aa} is the number of new group examinees that received an old form equivalent AP grade of *a* (where *a* ranges 1 to 5) using both the studied and criterion methods, and *N* is the total number of examinees in the new group. In this sense, the CC statistic is weighted by the relative frequency of the new form group.

Results

In this section, results are presented first according to the evaluation criteria and next by AP subject. Results of wRMSD for German, Italian, and French are presented in Tables 12 - 24. DTM results are presented in Tables 25-37, and CC results can be found in Tables 38-44. Due to the high volume of results, summary tables of wRMSDs, DTM proportions, and CC results can be found in Tables 24, 37, and 44, respectively. These summary tables collapse across equating methods and equating years in order to provide a clearer picture of the general performance of each studied anchor set.

Trends Organized by Criteria

wRMSD for raw composite scores and scale scores. In general, MC-only linking conditions resulted in the smallest wRMSD values for German and Italian whereas the MC + Best FR condition produced the smallest values for French. Specific trends in wRMSD values across all studied conditions can be found in Tables 12-23 and aggregated wRMSD trends can be found in Table 24. Across subjects, when the group difference was 0.3 and traditional equipercentile equating was used, the lowest wRMSD values were generally associated with conditions that used the MC + Best FR item to conduct scale linking.

The general pattern of results were similar across IRT and traditional methods with the exception of German and French when group differences reached 0.3. For these subjects, when IRT was used, the lowest wRMSD values tended to be associated with the MC-only anchor.

In general, the largest wRMSD values were found when the anchor set included all FR items. Several exceptions were found for French. First, when IRT was used, the largest wRMSDs were generally found when the anchor set included the worst FR item. When traditional methods were used, the largest wRMSD values were associated with the MC-only anchor conditions when group differences were largest (i.e., *ES* = 0.3).

The aforementioned trends were virtually identical for unrounded and rounded composite scores and unrounded and rounded scale scores except for in a few French conditions (Tables 12-23). For this reason, only results for unrounded scores were aggregated (Table 24).

Difference That Matters for unrounded raw scores and scale scores. The highest proportion of score points that fell within the DTM boundaries was generally associated with the MC-only and MC + Best FR anchor conditions. This was found across equating methods and at all levels of group differences (see Tables 25-37). When the *ES* reached 0.3, the MC + Best FR anchor set appeared to outperform the MC-only anchor conditions, but for German and French only.

There were no clear patterns that highlighted which anchor set performed most differently from the criterion with respect to the DTM criteria. In general, MC + Worst FR and MC + All FR anchor conditions resulted in the lowest proportion of equated raw scores that fell within the DTM boundaries. When IRT equating was used, the worst performing anchor set consisted of the MC + Worst FR item. This was found across all subjects and all levels of group differences. When traditional CHE and FEE methods were used, the anchor conditions with the MC + Worst FR and MC + All FR anchor sets resulted in the lowest proportion of equated score points that fell within the DTM boundary (Tables 25-36).

Classification consistency of AP grades. In general, the MC-only anchor conditions resulted in the highest agreement percentages across both IRT and traditional equating methods for German and Italian. For French, linking using the MC + Best FR item tended to result in the highest agreement, across all *ES* conditions and all equating methods (see Tables 38-44).

In general, the lowest agreement percentages were found in conditions that used MC + All FR items or MC + Worst FR item as the anchor set. This pattern was generally found for all three levels of group differences and across all equating methods. However, for French, the lowest agreement was often seen when the MC-only anchor was used, especially for the IRT methods.

Trends Organized by Subject

In this section, results are broken down by AP exam rather than by the type of criteria.

German. The classification consistency, wRMSD, and DTM results generally indicated that the MC-only anchor conditions performed closest to the criterion equating relationship (Tables 24, 37, 44). However there were a few exceptions; for example, when group differences were large, DTM and wRMSD results tended to favor the anchor set made up of the MC + Best FR item (Tables 12-15; 25-28; 38-49).

When IRT equating methods were used, the lowest classification agreement percentages were exclusively associated with the MC + All FR condition whereas when traditional equating methods were used, the lowest agreement was associated with a mixture of MC + All FR and MC + Worst FR conditions. The DTM results indicated that the MC + Worst FR conditions tended to result in equated scores that were most different from the criterion whereas the wRMSD results indicated the MC + All FR conditions performed most differently.

Italian. The classification consistency, wRMSD, and DTM results generally indicated that the MC-only anchor conditions performed closest to the criterion equating relationship for all equating methods and all levels of group differences (see Tables 24, 37, 44).

When IRT equating was used, the lowest CC percentages were associated with the MC + All FR conditions (Tables 40-41). For traditional equating methods, the MC + All FR and MC + Worst FR conditions led to the lowest agreement. The largest wRMSD values were associated with the MC + All FR conditions for all equating methods and levels of group differences (Tables 16-19). The DTM results generally indicated that the MC + All FR and MC + Worst FR conditions performed most differently from the criterion for the traditional and IRT equating methods, respectively (Tables 29-32).

In general, the MC + All FR condition performed most differently from the equating criterion for Italian. This finding agrees with the flagging results of the original study where FR items 1, 2, and 4 all performed relatively poor in comparison to FR 3 (best performing FR).

French. The CC, wRMSD, and DTM results generally indicated that the MC+ Best FR anchor condition performed closest to the criterion equating relationship using both IRT and traditional equating methods (Tables 20-23; 33-36; 42-43). When there were no group differences, the MC-only and MC + Best FR conditions performed similarly and were closest to

the criterion. When group differences existed, the MC + Best FR anchor condition generally performed closest to the criterion (Tables 24, 37, 44).

The lowest agreement percentages and largest wRMSD values were found for the MC-only and MC + Worst FR conditions when IRT was used and for the MC + All FR and MC-only CI conditions when traditional methods were used. When group differences were large (*ES* = 0.3), the MC-only conditions resulted in the lowest agreement percentages and the largest wRMSD values. The DTM results indicated that the MC + Worst FR and MC + All FR conditions resulted in equated scores that were least similar to the criterion and this was found across all equating methods and levels of group differences. One caveat worth mentioning is that when group differences were large and traditional equating methods were used, the MC-only condition resulted in the lowest within DTM boundary percentage for raw composite scores.

Discussion

The primary goal of this study was to investigate whether task model generated FR items could be used to improve equating results. In this study, the four anchor compositions studied included MC-only items, MC + the Best FR, MC + the Worst FR and MC + All FR items. Three levels of group difference effect sizes were considered (0, 0.1, 0.3) and were fully crossed with four equating methods, including FE, CE, IRT TS, IRT OS. The criterion equating relationships were established using the RG design, which was approximated using pseudo-group samples from the operational forms.

In general, the findings from the current study were mixed and did not routinely favor anchor sets that included task model derived FR item(s). More specifically, unequivocal evidence to support the use of FR items in the anchor set was not found for any of the subjects. However, for French tests, the results were more favorable in the sense that the MC + Best FR condition typically showed smaller wRMSDs and larger CC and DTM proportions compared to any other studied anchor condition. For German and Italian, the MC-only condition still performed better than other three anchor conditions. Since the same AP task model rubrics were applied to all three subjects, it is unclear at this time why results for French were quite different from those of German and Italian. One potential explanation is that the French MC and FR items might measure somewhat more distinct constructs compared to the German and Italian tests. This hypothesis is somewhat supported by the lower disattenuated correlation between MC and FR section scores for French. However, additional research is needed to determine whether this

hypothesis is better supported. When group differences were large (*ES* = 0.3), including the best performing FR item in the anchor set tended to improve equating results. This finding is consistent with previous research (Wang & Kolen, 2014).

According to previous research (Luecht, 2013), items generated from the same task model share similar psychometric properties and possibly perform as isomorphs from a statistical or psychometric perspective. This study investigated the extent to which equating results are affected by using task model derived FR items as part of the anchor set. Unfortunately, the findings from the current study did not entirely support the notion that task model items could or should be used interchangeably in equating contexts. Findings from this study cannot necessarily be generalized to other testing programs that use task models for item development. This lack of generalizability is present because the design and depth of task model rubrics can potentially vary considerably which can therefore lead to items that are more or less similar to one another.

It is important to point out that this study relied on results from Malatesta and Liu (2016) in order to label which task model derived FR items might be best suited for equating purposes. In their study, methods that were originally developed for MC anchor items were adapted to identify good and poor performing FR items. The fact that these methods were originally designed for MC items may have introduced additional sources of error into the current study. Furthermore, as Malatesta and Liu (2016) pointed out, task models and FR items were entirely confounded within each exam. Furthermore, examinees were only exposed to one item from each task model. While this data collection design reflected the operational administration, it also prohibited the use of more complex, yet perhaps more informative analyses (i.e., generalizability theory analyses or hierarchical IRT modeling) from being conducted. In the future, a special study, with a different data collection design would be beneficial in order to more thoroughly evaluate the degree of similarity between items developed from the same task model.

References

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 199–218). Mahwah, NJ: Lawrence Erlbaum Associates.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph Number 1). Iowa City, IA: CASMA, The University of Iowa.

Cai, L. (2015). *flexMIRT* (Version 3.0) [Computer Program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

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.

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

Hagge, S. L. (2010). *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.

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

Kim, S. & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* (Version 1.0) [Computer Program]. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Kim, S., & Kolen, M. J. (2004a). *Optimally defining criterion functions for the characteristic curve procedures in the IRT scale linking*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43,* 53-76.

Luecht, R. M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D.J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Luecht, R. M. (2013). Assessment engineering task model maps: task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14,* 1-38.

Malatesta, J., & Liu, H. (2016). Evaluating the interchangeability of free-response items developed from task models. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 4). (CASMA Monograph No. 2.4).* Iowa City, IA: CASMA, The University of Iowa.

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

Wang, W., & Kolen, M. J. (2014). Comparison of the use of mc only and mixed-format common items in mixed-format test score equating. 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: CASMA, The University of Iowa.

Table 1

Descriptive Statistics for French Link 12-13 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	51.8	14.9	-0.19	-0.65	1	84	3000
		MC-only CI	13.6	4.59	-0.11	-0.83	0	23	
		MC + best FR CI	16.5	5.24	-0.11	-0.76	0	28	
		MC + worst FR CI	15.9	5.48	-0.07	-0.80	1	28	
		MC + all FR CI	24.4	8.06	-0.14	-0.68	1	43	
2013	0	Total Score	51.7	14.8	-0.21	-0.70	11	84	3000
		MC-only CI	13.6	4.59	-0.08	-0.86	2	23	
		MC + best FR CI	16.5	5.25	-0.07	-0.82	3	28	
		MC + worst FR CI	16.4	5.40	-0.09	-0.78	3	28	
		MC + all FR CI	25.2	7.85	-0.15	-0.64	4	43	
2012	0.1	Total Score	49.7	14.9	-0.06	-0.77	10	84	3000
		MC-only CI	12.9	4.56	0.01	-0.86	2	23	
		MC + best FR CI	15.7	5.21	0.02	-0.84	2	28	
		MC + worst FR CI	15.1	5.43	0.05	-0.83	2	28	
		MC + all FR CI	23.2	8.04	0.00	-0.76	3	43	
2013	0.1	Total Score	51.4	15.1	-0.20	-0.74	11	84	3000
		MC-only CI	13.5	4.64	-0.05	-0.85	0	23	
		MC + best FR CI	16.3	5.31	-0.03	-0.83	3	28	
		MC + worst FR CI	16.3	5.46	-0.07	-0.80	3	28	
		MC + all FR CI	25.0	7.96	-0.13	-0.69	4	43	
2012	0.3	Total Score	48.8	14.9	-0.03	-0.72	11	84	3000
		MC-only CI	12.7	4.54	0.06	-0.80	0	23	
		MC + best FR CI	15.4	5.17	0.07	-0.79	2	28	
		MC + worst FR CI	14.8	5.41	0.11	-0.77	2	28	
		MC + all FR CI	22.7	7.99	0.05	-0.72	2	43	
2013	0.3	Total Score	53.7	14.4	-0.33	-0.54	13	84	3000
		MC-only CI	14.1	4.50	-0.18	-0.76	2	23	
		MC + best FR CI	17.1	5.13	-0.17	-0.72	3	28	
		MC + worst FR CI	17.1	5.25	-0.20	-0.65	2	28	
		MC + all FR CI	26.3	7.63	-0.27	-0.45	4	43	

Table 2

Descriptive Statistics for French Link 12-14 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	51.5	14.8	-0.15	-0.67	8	84	3000
		MC-only CI	13.9	4.16	-0.23	-0.53	0	22	
		MC + best FR CI	16.8	4.80	-0.20	-0.53	2	27	
		MC + worst FR CI	16.2	4.95	-0.18	-0.49	2	27	
		MC + all FR CI	24.6	7.55	-0.18	-0.48	3	42	
2014	0	Total Score	51.3	14.3	-0.14	-0.51	5	84	3000
		MC-only CI	13.9	4.11	-0.25	-0.43	0	22	
		MC + best FR CI	16.7	4.87	-0.17	-0.49	1	27	
		MC + worst FR CI	16.7	4.89	-0.20	-0.41	1	27	
		MC + all FR CI	25.4	7.47	-0.14	-0.48	3	42	
2012	0.1	Total Score	49.5	14.8	-0.04	-0.70	8	84	3000
		MC-only CI	13.4	4.20	-0.14	-0.54	1	22	
		MC + best FR CI	16.1	4.83	-0.11	-0.52	2	27	
		MC + worst FR CI	15.6	5.03	-0.08	-0.53	1	27	
		MC + all FR CI	23.5	7.65	-0.08	-0.50	2	42	
2014	0.1	Total Score	51.1	14.3	-0.13	-0.53	5	85	3000
		MC-only CI	13.8	4.09	-0.22	-0.44	0	22	
		MC + best FR CI	16.6	4.86	-0.16	-0.48	1	27	
		MC + worst FR CI	16.7	4.89	-0.17	-0.45	1	27	
		MC + all FR CI	25.3	7.48	-0.12	-0.49	4	42	
2012	0.3	Total Score	50.3	15.0	-0.10	-0.76	10	84	3000
		MC-only CI	13.6	4.18	-0.16	-0.59	0	22	
		MC + best FR CI	16.4	4.83	-0.12	-0.58	3	27	
		MC + worst FR CI	15.8	5.03	-0.09	-0.61	2	27	
		MC + all FR CI	24.0	7.70	-0.12	-0.58	3	42	
2014	0.3	Total Score	55.6	13.8	-0.30	-0.40	10	85	3000
		MC-only CI	14.9	3.93	-0.37	-0.32	0	22	
		MC + best FR CI	18.0	4.67	-0.32	-0.39	2	27	
		MC + worst FR CI	18.0	4.70	-0.31	-0.34	3	27	
		MC + all FR CI	27.5	7.17	-0.28	-0.38	4	42	

Table 3

Descriptive Statistics for French Link 13-15 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2013	0	Total Score	49.7	14.9	-0.10	-0.79	12	85	3000
		MC-only CI	14.6	4.82	-0.22	-0.73	2	24	
		MC + best FR CI	17.4	5.51	-0.18	-0.72	2	29	
		MC + worst FR CI	17.3	5.66	-0.18	-0.71	2	29	
		MC + all FR CI	25.6	8.20	-0.19	-0.62	3	44	
2015	0	Total Score	46.7	14.3	0.17	-0.58	10	83	3000
		MC-only CI	14.6	4.66	-0.18	-0.66	1	24	
		MC + best FR CI	17.6	5.39	-0.14	-0.64	2	29	
		MC + worst FR CI	17.1	5.38	-0.14	-0.62	2	29	
		MC + all FR CI	25.5	7.93	-0.07	-0.55	3	44	
2013	0.1	Total Score	49.6	14.8	-0.08	-0.74	4	83	3000
		MC-only CI	14.5	4.76	-0.19	-0.70	2	24	
		MC + best FR CI	17.3	5.43	-0.15	-0.68	2	29	
		MC + worst FR CI	17.2	5.57	-0.16	-0.68	2	29	
		MC + all FR CI	25.6	8.07	-0.14	-0.61	2	44	
2015	0.1	Total Score	48.2	14.2	0.10	-0.65	10	84	3000
		MC-only CI	15.1	4.59	-0.23	-0.68	2	24	
		MC + best FR CI	18.1	5.32	-0.21	-0.66	2	29	
		MC + worst FR CI	17.6	5.29	-0.20	-0.61	2	29	
		MC + all FR CI	26.4	7.83	-0.15	-0.55	3	44	
2013	0.3	Total Score	48.8	14.9	-0.06	-0.79	4	83	3000
		MC-only CI	14.3	4.80	-0.18	-0.72	2	24	
		MC + best FR CI	17.0	5.48	-0.14	-0.70	2	29	
		MC + worst FR CI	16.9	5.66	-0.15	-0.71	2	29	
		MC + all FR CI	25.2	8.18	-0.14	-0.62	2	44	
2015	0.3	Total Score	50.9	14.0	0.01	-0.64	12	84	3000
		MC-only CI	15.9	4.43	-0.37	-0.47	1	24	
		MC + best FR CI	19.0	5.13	-0.34	-0.44	2	29	
		MC + worst FR CI	18.6	5.10	-0.34	-0.42	3	29	
		MC + all FR CI	27.9	7.53	-0.26	-0.39	4	44	

Table 4

Descriptive Statistics for German Link 12-13 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	57.7	16.4	-0.33	-0.72	8	85	1600
		MC-only CI	13.9	4.30	-0.25	-0.68	0	21	
		MC + best FR CI	17.5	5.24	-0.29	-0.60	2	26	
		MC + worst FR CI	16.9	5.38	-0.24	-0.71	2	26	
		MC + all FR CI	27.6	8.33	-0.37	-0.52	2	41	
2013	0	Total Score	58.4	15.3	-0.21	-0.69	15	85	2600
		MC-only CI	13.9	4.03	-0.19	-0.63	0	21	
		MC + best FR CI	17.4	4.86	-0.20	-0.56	3	26	
		MC + worst FR CI	17.2	5.05	-0.25	-0.60	2	26	
		MC + all FR CI	28.0	7.59	-0.29	-0.49	4	41	
2012	0.1	Total Score	57.8	16.5	-0.35	-0.73	8	85	1550
		MC-only CI	14.0	4.35	-0.26	-0.74	0	21	
		MC + best FR CI	17.5	5.30	-0.30	-0.67	3	26	
		MC + worst FR CI	16.9	5.41	-0.23	-0.77	3	26	
		MC + all FR CI	27.7	8.39	-0.37	-0.60	4	41	
2013	0.1	Total Score	56.7	15.7	-0.15	-0.73	10	85	2500
		MC-only CI	13.5	4.15	-0.14	-0.71	2	21	
		MC + best FR CI	16.8	5.01	-0.15	-0.62	2	26	
		MC + worst FR CI	16.7	5.22	-0.17	-0.72	2	26	
		MC + all FR CI	27.2	7.85	-0.23	-0.61	4	41	
2012	0.3	Total Score	58.0	16.4	-0.33	-0.72	8	85	1600
		MC-only CI	14.1	4.33	-0.28	-0.70	0	21	
		MC + best FR CI	17.6	5.25	-0.31	-0.64	2	26	
		MC + worst FR CI	17.0	5.40	-0.25	-0.75	2	26	
		MC + all FR CI	27.8	8.32	-0.37	-0.56	2	41	
2013	0.3	Total Score	53.3	15.8	0.04	-0.68	9	85	2700
		MC-only CI	12.8	4.12	0.04	-0.66	2	21	
		MC + best FR CI	15.9	5.04	0.02	-0.55	2	26	
		MC + worst FR CI	15.6	5.22	-0.02	-0.66	2	26	
		MC + all FR CI	25.6	7.99	-0.07	-0.59	3	41	

Table 5

Descriptive Statistics for German Link 12-14 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	56.6	16.8	-0.25	-0.83	8	85	1600
		MC-only CI	13.0	3.99	-0.39	-0.71	0	19	
		MC + best FR CI	16.5	4.94	-0.38	-0.65	1	24	
		MC + worst FR CI	15.8	5.09	-0.33	-0.71	1	24	
		MC + all FR CI	26.4	8.13	-0.39	-0.63	3	39	
2014	0	Total Score	60.8	15.4	-0.45	-0.52	5	85	2400
		MC-only CI	12.9	3.86	-0.36	-0.60	0	19	
		MC + best FR CI	16.7	4.61	-0.39	-0.50	0	24	
		MC + worst FR CI	16.4	4.83	-0.37	-0.62	0	24	
		MC + all FR CI	27.6	7.41	-0.44	-0.49	3	39	
2012	0.1	Total Score	57.7	16.6	-0.31	-0.78	11	85	1700
		MC-only CI	13.3	3.96	-0.45	-0.66	0	19	
		MC + best FR CI	16.8	4.89	-0.45	-0.59	2	24	
		MC + worst FR CI	16.2	5.06	-0.39	-0.67	1	24	
		MC + all FR CI	26.9	8.04	-0.45	-0.55	3	39	
2014	0.1	Total Score	60.7	15.3	-0.43	-0.53	5	85	2500
		MC-only CI	12.8	3.88	-0.35	-0.57	0	19	
		MC + best FR CI	16.6	4.62	-0.37	-0.50	0	24	
		MC + worst FR CI	16.3	4.83	-0.35	-0.61	0	24	
		MC + all FR CI	27.5	7.38	-0.41	-0.48	3	39	
2012	0.3	Total Score	57.7	16.5	-0.31	-0.79	11	85	1700
		MC-only CI	13.3	3.92	-0.43	-0.72	1	19	
		MC + best FR CI	16.8	4.85	-0.43	-0.63	2	24	
		MC + worst FR CI	16.2	5.02	-0.38	-0.69	1	24	
		MC + all FR CI	26.9	7.99	-0.45	-0.56	3	39	
2014	0.3	Total Score	56.3	16.8	-0.21	-0.82	5	85	2200
		MC-only CI	12.0	4.03	-0.11	-0.82	0	19	
		MC + best FR CI	15.5	4.89	-0.14	-0.77	0	24	
		MC + worst FR CI	15.2	5.11	-0.13	-0.83	0	24	
		MC + all FR CI	25.6	8.03	-0.21	-0.74	2	39	

Table 6

Descriptive Statistics for German Link 13-15 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2013	0	Total Score	54.2	15.8	-0.02	-0.70	9	85	2400
		MC-only CI	14.8	4.96	0.08	-0.79	0	24	
		MC + best FR CI	18.0	5.85	0.05	-0.72	1	29	
		MC + worst FR CI	17.8	6.07	0.01	-0.75	0	29	
		MC + all FR CI	27.9	8.81	-0.07	-0.66	3	44	
2015	0	Total Score	52.6	16.9	0.07	-0.83	9	85	2400
		MC-only CI	14.8	4.93	0.12	-0.76	0	24	
		MC + best FR CI	18.0	5.81	0.10	-0.70	2	29	
		MC + worst FR CI	18.0	5.94	0.06	-0.74	0	29	
		MC + all FR CI	28.1	8.77	-0.04	-0.65	2	44	
2013	0.1	Total Score	55.9	16.3	-0.14	-0.77	9	85	2600
		MC-only CI	15.4	5.09	-0.05	-0.83	0	24	
		MC + best FR CI	18.6	6.02	-0.08	-0.77	1	29	
		MC + worst FR CI	18.4	6.22	-0.10	-0.84	0	29	
		MC + all FR CI	28.7	9.05	-0.17	-0.72	3	44	
2015	0.1	Total Score	52.5	16.9	0.09	-0.85	9	85	2400
		MC-only CI	14.8	4.96	0.12	-0.75	0	24	
		MC + best FR CI	18.0	5.83	0.09	-0.70	1	29	
		MC + worst FR CI	18.0	5.95	0.07	-0.74	0	29	
		MC + all FR CI	28.0	8.78	-0.04	-0.64	2	44	
2013	0.3	Total Score	58.2	15.7	-0.29	-0.61	9	85	2400
		MC-only CI	16.0	4.98	-0.18	-0.78	0	24	
		MC + best FR CI	19.4	5.85	-0.22	-0.67	3	29	
		MC + worst FR CI	19.3	6.05	-0.25	-0.72	3	29	
		MC + all FR CI	30.0	8.69	-0.33	-0.53	4	44	
2015	0.3	Total Score	51.5	16.7	0.14	-0.80	9	85	2400
		MC-only CI	14.5	4.87	0.17	-0.71	0	24	
		MC + best FR CI	17.7	5.74	0.15	-0.66	2	29	
		MC + worst FR CI	17.7	5.86	0.11	-0.69	0	29	
		MC + all FR CI	27.6	8.69	0.01	-0.62	5	44	

Table 7

Descriptive Statistics for Italian Link 12-13 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	58.1	16.8	-0.16	-0.61	13	90	1200
		MC-only CI	13.6	4.17	-0.36	-0.62	2	20	
		MC + best FR CI	16.9	5.26	-0.32	-0.66	2	25	
		MC + worst FR CI	16.4	5.18	-0.29	-0.61	2	25	
		MC + all FR CI	25.0	8.27	-0.25	-0.53	2	40	
2013	0	Total Score	58.3	18.1	-0.10	-0.95	13	90	1200
		MC-only CI	13.6	4.21	-0.33	-0.68	0	20	
		MC + best FR CI	16.9	5.32	-0.30	-0.77	3	25	
		MC + worst FR CI	15.7	5.61	-0.13	-0.92	2	25	
		MC + all FR CI	26.0	8.66	-0.23	-0.87	3	40	
2012	0.1	Total Score	58.9	16.9	-0.14	-0.73	13	90	1300
		MC-only CI	13.8	4.18	-0.37	-0.61	2	20	
		MC + best FR CI	17.0	5.25	-0.31	-0.72	2	25	
		MC + worst FR CI	16.6	5.17	-0.30	-0.61	2	25	
		MC + all FR CI	25.4	8.27	-0.23	-0.60	2	40	
2013	0.1	Total Score	56.5	17.8	-0.03	-0.93	13	90	1300
		MC-only CI	13.3	4.11	-0.29	-0.56	0	20	
		MC + best FR CI	16.5	5.16	-0.22	-0.74	3	25	
		MC + worst FR CI	15.2	5.44	-0.05	-0.89	2	25	
		MC + all FR CI	25.2	8.46	-0.14	-0.85	3	40	
2012	0.3	Total Score	60.4	16.8	-0.24	-0.66	13	90	1000
		MC-only CI	14.1	4.17	-0.47	-0.56	2	20	
		MC + best FR CI	17.5	5.25	-0.43	-0.62	3	25	
		MC + worst FR CI	17.0	5.15	-0.40	-0.56	2	25	
		MC + all FR CI	26.1	8.25	-0.34	-0.54	3	40	
2013	0.3	Total Score	54.9	17.5	0.00	-0.90	13	90	1000
		MC-only CI	12.9	4.13	-0.24	-0.60	0	20	
		MC + best FR CI	16.0	5.16	-0.18	-0.76	3	25	
		MC + worst FR CI	14.7	5.40	0.00	-0.88	1	25	
		MC + all FR CI	24.3	8.43	-0.10	-0.84	3	40	

Table 8

Descriptive Statistics for Italian Link 12-14 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2012	0	Total Score	57.5	16.6	-0.14	-0.65	13	90	1200
		MC-only CI	13.0	3.59	-0.13	-0.56	3	20	
		MC + best FR CI	16.2	4.64	-0.14	-0.62	3	25	
		MC + worst FR CI	15.7	4.52	-0.10	-0.55	3	25	
		MC + all FR CI	24.2	7.60	-0.11	-0.45	3	40	
2014	0	Total Score	53.2	17.1	0.06	-0.83	9	88	1200
		MC-only CI	13.0	3.70	-0.22	-0.40	0	20	
		MC + best FR CI	16.4	4.79	-0.21	-0.60	2	25	
		MC + worst FR CI	16.0	4.75	-0.24	-0.38	0	25	
		MC + all FR CI	25.5	8.17	-0.25	-0.62	3	40	
2012	0.1	Total Score	59.0	17.1	-0.20	-0.69	13	90	1300
		MC-only CI	13.3	3.66	-0.21	-0.53	3	20	
		MC + best FR CI	16.6	4.73	-0.23	-0.56	3	25	
		MC + worst FR CI	16.1	4.64	-0.17	-0.54	4	25	
		MC + all FR CI	24.9	7.82	-0.18	-0.47	4	40	
2014	0.1	Total Score	52.6	17.3	0.11	-0.88	9	88	1300
		MC-only CI	12.9	3.74	-0.19	-0.48	0	20	
		MC + best FR CI	16.2	4.85	-0.17	-0.64	2	25	
		MC + worst FR CI	15.8	4.81	-0.20	-0.47	0	25	
		MC + all FR CI	25.2	8.28	-0.20	-0.67	3	40	
2012	0.3	Total Score	59.4	17.1	-0.19	-0.70	13	90	1200
		MC-only CI	13.5	3.67	-0.21	-0.57	3	20	
		MC + best FR CI	16.7	4.76	-0.21	-0.64	3	25	
		MC + worst FR CI	16.3	4.66	-0.18	-0.58	4	25	
		MC + all FR CI	25.2	7.83	-0.17	-0.54	4	40	
2014	0.3	Total Score	49.8	16.6	0.23	-0.63	8	88	1200
		MC-only CI	12.3	3.63	-0.06	-0.37	0	20	
		MC + best FR CI	15.5	4.69	-0.06	-0.53	2	25	
		MC + worst FR CI	15.1	4.63	-0.05	-0.36	2	25	
		MC + all FR CI	23.8	8.01	-0.07	-0.59	4	40	

Table 9

Descriptive Statistics for Italian Link 13-15 Pseudo Group Data under CINEG Design

Form	ES	Score Description	Mean	SD	Skew	Kurt	Min	Max	Sample
2013	0	Total Score	56.7	17.9	-0.01	-0.96	13	90	1300
		MC-only CI	12.8	4.45	-0.20	-0.92	0	20	
		MC + best FR CI	16.1	5.51	-0.18	-0.99	3	25	
		MC + worst FR CI	14.8	5.78	-0.02	-1.07	2	25	
		MC + all FR CI	24.8	8.78	-0.09	-0.99	4	40	
2015	0	Total Score	52.5	18.3	0.16	-0.94	5	90	1300
		MC-only CI	12.8	4.47	-0.15	-0.92	0	20	
		MC + best FR CI	16.2	5.53	-0.15	-0.95	2	25	
		MC + worst FR CI	15.0	5.61	-0.01	-0.89	0	25	
		MC + all FR CI	24.3	8.87	-0.03	-0.86	3	40	
2013	0.1	Total Score	57.7	18.2	-0.07	-0.96	13	90	1400
		MC-only CI	13.1	4.47	-0.25	-0.87	0	20	
		MC + best FR CI	16.3	5.57	-0.21	-0.96	3	25	
		MC + worst FR CI	15.1	5.84	-0.06	-1.05	2	25	
		MC + all FR CI	25.3	8.90	-0.14	-0.99	4	40	
2015	0.1	Total Score	51.7	18.6	0.20	-0.95	5	90	1400
		MC-only CI	12.6	4.50	-0.10	-0.94	0	20	
		MC + best FR CI	15.9	5.60	-0.10	-0.98	2	25	
		MC + worst FR CI	14.7	5.66	0.04	-0.91	0	25	
		MC + all FR CI	23.9	9.00	0.01	-0.89	3	40	
2013	0.3	Total Score	58.9	18.4	-0.13	-0.98	13	90	1000
		MC-only CI	13.4	4.48	-0.30	-0.91	0	20	
		MC + best FR CI	16.7	5.61	-0.27	-0.95	3	25	
		MC + worst FR CI	15.5	5.89	-0.13	-1.07	2	25	
		MC + all FR CI	25.8	8.97	-0.20	-0.99	4	40	
2015	0.3	Total Score	49.5	17.5	0.25	-0.77	14	89	1000
		MC-only CI	12.2	4.36	-0.12	-0.90	1	20	
		MC + best FR CI	15.4	5.48	-0.11	-0.95	2	25	
		MC + worst FR CI	14.1	5.37	0.04	-0.83	1	25	
		MC + all FR CI	22.8	8.62	0.01	-0.81	3	40	

Table 10

Additional Exam Characteristics for Pseudo Group CINEG Data

Data Characteristic	French	German	Italian
Average Disattenuated Correlation (MC,FR sections)	0.89	0.93	0.95
Best FR Item	FR 1	FR 2	FR 3
Worst FR Item	FR 4	FR 4	FR 4

Table 11

Descriptive Statistics for French, German and Italian under Random Groups Design

Subject	Link	Form	Mean	SD	Skew	Kurt	Min	Max	Sample
French	12-13	2012	55.9	14.6	-0.40	-0.51	1	85	13000
		2013	56.1	14.6	-0.44	-0.51	0	85	13000
	12-14	2012	55.8	14.6	-0.39	-0.48	1	85	13000
		2014	55.3	14.2	-0.29	-0.45	5	85	13000
	13-15	2013	55.4	14.7	-0.41	-0.55	0	85	14500
		2015	52.1	14.3	-0.09	-0.67	9	84	14500
German	12-13	2012	57.0	16.6	-0.27	-0.80	8	85	1840
		2013	58.0	15.3	-0.20	-0.71	15	85	3100
	12-14	2012	57.1	16.7	-0.29	-0.78	8	85	1840
		2014	61.6	15.2	-0.51	-0.41	5	85	2050
	13-15	2013	55.7	15.8	-0.10	-0.75	9	85	4100
		2015	54.2	16.8	0.00	-0.88	9	85	4300
Italian	12-13	2012	46.0	12.4	-0.19	-0.68	11	69	1405
		2013	45.1	13.8	-0.09	-0.89	0	69	1300
	12-14	2012	46.1	12.4	-0.19	-0.70	11	69	1390
		2014	40.8	12.9	0.22	-0.80	0	69	1450
	13-15	2013	44.3	13.5	-0.06	-0.82	0	69	1740
		2015	40.8	13.8	0.13	-0.97	0	69	1900

Table 12

Weighted Root Mean Squared Differences of Unrounded Composite Scores for German Using IRT Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	2.824	2.419	3.401	3.316	2.990
IRT TS	2.911	2.507	3.469	3.411	3.075
2012-2014					
IRT OS	1.816	3.964	2.746	5.062	3.397
IRT TS	1.840	4.004	2.759	5.080	3.421
2013-2015					
IRT OS	0.904	0.687	1.131	0.261	0.746
IRT TS	0.922	0.707	1.156	0.268	0.763
Average	1.870	2.381	2.444	2.900	2.399
			ES = 0.1		
2012-2013					
IRT OS	2.336	2.376	2.890	3.132	2.683
IRT TS	2.396	2.445	2.936	3.206	2.746
2012-2014					
IRT OS	1.258	3.496	2.184	4.857	2.949
IRT TS	1.292	3.545	2.211	4.882	2.983
2013-2015					
IRT OS	0.577	0.281	0.791	0.327	0.494
IRT TS	0.583	0.297	0.815	0.325	0.505
Average	1.407	2.073	1.971	2.788	2.060
			ES = 0.3		
2012-2013					
IRT OS	2.553	1.876	3.397	3.281	2.777
IRT TS	2.615	1.941	3.447	3.341	2.836
2012-2014					
IRT OS	2.060	4.409	3.331	5.524	3.831
IRT TS	2.060	4.408	3.324	5.511	3.826
2013-2015					
IRT OS	0.992	0.726	1.632	0.618	0.992
IRT TS	1.025	0.749	1.663	0.618	1.014
Average	1.884	2.351	2.799	3.149	2.546

Table 13

Weighted Root Mean Squared Differences of Unrounded Scale Scores for German Using IRT Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	1.272	1.255	1.361	1.444	1.333
IRT TS	1.227	1.177	1.341	1.403	1.287
2012-2014					
IRT OS	0.778	1.499	1.022	1.828	1.282
IRT TS	0.749	1.490	1.013	1.823	1.269
2013-2015					
IRT OS	0.462	0.363	0.547	0.115	0.372
IRT TS	0.436	0.346	0.523	0.121	0.356
Average	0.821	1.022	0.968	1.122	0.983
			ES = 0.1		
2012-2013					
IRT OS	1.124	1.299	1.207	1.413	1.261
IRT TS	1.058	1.198	1.168	1.354	1.195
2012-2014					
IRT OS	0.608	1.324	0.830	1.738	1.125
IRT TS	0.572	1.318	0.821	1.737	1.112
2013-2015					
IRT OS	0.333	0.242	0.452	0.156	0.296
IRT TS	0.293	0.204	0.407	0.139	0.261
Average	0.665	0.931	0.814	1.089	0.875
			ES = 0.3		
2012-2013					
IRT OS	1.229	1.073	1.441	1.443	1.297
IRT TS	1.159	0.978	1.391	1.385	1.228
2012-2014					
IRT OS	0.828	1.630	1.232	1.980	1.418
IRT TS	0.801	1.613	1.215	1.966	1.399
2013-2015					
IRT OS	0.570	0.413	0.780	0.244	0.502
IRT TS	0.513	0.372	0.727	0.233	0.461
Average	0.850	1.013	1.131	1.209	1.051

Table 14

Weighted Root Mean Squared Differences of Unrounded Composite Scores for German Using Traditional Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	2.067	2.161	2.985	3.140	2.588
FEE	1.569	1.773	2.741	3.049	2.283
2012-2014					
CHE	1.472	2.876	3.942	5.394	3.421
FEE	1.064	2.528	3.649	5.284	3.131
2013-2015					
CHE	2.505	2.835	4.356	4.127	3.456
FEE	1.876	2.221	3.929	3.862	2.972
Average	1.759	2.399	3.600	4.143	2.975
			ES = 0.1		
2012-2013					
CHE	2.109	2.198	2.930	3.061	2.574
FEE	1.864	1.897	2.880	3.013	2.413
2012-2014					
CHE	1.440	2.568	3.600	5.266	3.218
FEE	0.920	2.282	3.431	5.181	2.953
2013-2015					
CHE	2.353	2.625	4.271	4.157	3.352
FEE	2.066	2.301	4.193	4.125	3.171
Average	1.792	2.312	3.551	4.134	2.947
			ES = 0.3		
2012-2013					
CHE	2.947	2.612	3.528	3.216	3.076
FEE	3.083	2.533	3.704	3.243	3.141
2012-2014					
CHE	2.809	3.753	4.983	5.829	4.344
FEE	3.250	4.020	5.260	5.883	4.603
2013-2015					
CHE	3.969	3.497	5.316	4.345	4.282
FEE	4.652	3.752	5.823	4.403	4.657
Average	3.452	3.361	4.769	4.487	4.017

Table 15

Weighted Root Mean Squared Differences of Unrounded Scale Scores for German Using Traditional Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.900	0.974	1.115	1.230	1.055
FEE	0.663	0.781	1.002	1.184	0.907
2012-2014					
CHE	0.607	1.106	1.479	2.002	1.299
FEE	0.437	0.967	1.351	1.958	1.178
2013-2015					
CHE	1.114	1.180	1.633	1.538	1.366
FEE	0.787	0.958	1.474	1.459	1.170
Average	0.751	0.994	1.342	1.562	1.162
			ES = 0.1		
2012-2013					
CHE	0.902	0.943	1.086	1.192	1.031
FEE	0.713	0.779	1.036	1.159	0.922
2012-2014					
CHE	0.585	0.995	1.346	1.955	1.220
FEE	0.366	0.870	1.268	1.923	1.107
2013-2015					
CHE	0.904	1.000	1.547	1.516	1.242
FEE	0.752	0.849	1.519	1.499	1.155
Average	0.704	0.906	1.300	1.541	1.113
			ES = 0.3		
2012-2013					
CHE	1.188	1.058	1.292	1.244	1.196
FEE	1.149	0.983	1.338	1.238	1.177
2012-2014					
CHE	1.092	1.442	1.853	2.160	1.637
FEE	1.238	1.521	1.943	2.180	1.720
2013-2015					
CHE	1.488	1.332	1.928	1.610	1.590
FEE	1.676	1.380	2.091	1.619	1.691
Average	1.305	1.286	1.741	1.675	1.502

Table 16

Weighted Root Mean Squared Differences of Unrounded Composite Scores for Italian Using IRT Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.943	1.164	2.468	4.476	2.263
IRT TS	0.956	1.177	2.495	4.474	2.276
2012-2014					
IRT OS	2.867	4.360	4.203	7.558	4.747
IRT TS	2.861	4.351	4.194	7.547	4.738
2013-2015					
IRT OS	1.072	1.989	2.577	0.531	1.542
IRT TS	1.063	1.975	2.565	0.530	1.533
Average	1.627	2.503	3.084	4.186	2.850
			ES = 0.1		
2012-2013					
IRT OS	1.115	1.392	1.812	4.847	2.291
IRT TS	1.107	1.384	1.834	4.826	2.288
2012-2014					
IRT OS	2.889	4.341	4.142	7.655	4.756
IRT TS	2.892	4.339	4.144	7.649	4.756
2013-2015					
IRT OS	0.804	1.769	2.226	0.924	1.431
IRT TS	0.796	1.766	2.237	0.898	1.424
Average	1.600	2.498	2.733	4.466	2.824
			ES = 0.3		
2012-2013					
IRT OS	0.963	1.448	2.042	4.539	2.248
IRT TS	0.950	1.423	2.047	4.504	2.231
2012-2014					
IRT OS	2.934	4.538	3.874	7.270	4.654
IRT TS	2.928	4.509	3.862	7.256	4.639
2013-2015					
IRT OS	2.197	3.451	3.070	0.829	2.387
IRT TS	2.196	3.445	3.064	0.800	2.376
Average	2.028	3.136	2.993	4.200	3.089

Table 17

Weighted Root Mean Squared Differences of Unrounded Scale Scores for Italian Using IRT Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.538	0.637	1.058	1.984	1.054
IRT TS	0.501	0.595	1.030	1.886	1.003
2012-2014					
IRT OS	1.261	1.865	1.804	3.229	2.040
IRT TS	1.216	1.803	1.743	3.121	1.971
2013-2015					
IRT OS	0.401	0.744	0.971	0.236	0.588
IRT TS	0.399	0.737	0.962	0.231	0.582
Average	0.719	1.064	1.261	1.781	1.206
			ES = 0.1		
2012-2013					
IRT OS	0.524	0.658	0.792	2.074	1.012
IRT TS	0.498	0.621	0.777	1.977	0.968
2012-2014					
IRT OS	1.256	1.846	1.776	3.256	2.033
IRT TS	1.218	1.791	1.722	3.155	1.972
2013-2015					
IRT OS	0.343	0.692	0.842	0.360	0.559
IRT TS	0.332	0.681	0.841	0.352	0.551
Average	0.695	1.048	1.125	1.862	1.183
			ES = 0.3		
2012-2013					
IRT OS	0.488	0.714	0.860	2.024	1.021
IRT TS	0.459	0.663	0.846	1.900	0.967
2012-2014					
IRT OS	1.228	1.826	1.603	3.027	1.921
IRT TS	1.198	1.779	1.559	2.937	1.868
2013-2015					
IRT OS	0.948	1.420	1.172	0.384	0.981
IRT TS	0.915	1.379	1.155	0.371	0.955
Average	0.873	1.297	1.199	1.774	1.286

Table 18

Weighted Root Mean Squared Differences of Unrounded Composite Scores for Italian Using Traditional Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	1.414	1.529	5.149	4.052	3.036
FEE	1.258	1.341	4.697	3.887	2.796
2012-2014					
CHE	1.266	1.484	2.487	4.706	2.486
FEE	1.022	1.291	2.074	4.568	2.239
2013-2015					
CHE	1.831	2.279	3.584	3.335	2.757
FEE	1.390	1.861	3.076	3.284	2.403
Average	1.364	1.631	3.511	3.972	2.619
			ES = 0.1		
2012-2013					
CHE	1.354	1.457	4.695	4.141	2.912
FEE	1.633	1.653	4.195	4.085	2.891
2012-2014					
CHE	1.392	1.719	2.396	4.806	2.578
FEE	1.148	1.660	2.265	4.733	2.452
2013-2015					
CHE	1.975	2.508	2.967	3.212	2.666
FEE	2.267	2.729	2.573	3.201	2.692
Average	1.628	1.954	3.182	4.030	2.698
			ES = 0.3		
2012-2013					
CHE	1.692	1.740	4.837	4.182	3.113
FEE	2.443	2.345	4.299	4.214	3.325
2012-2014					
CHE	1.794	2.002	2.304	4.499	2.650
FEE	2.210	2.530	2.624	4.564	2.982
2013-2015					
CHE	2.533	3.496	3.159	3.549	3.184
FEE	4.602	4.943	3.304	3.751	4.150
Average	2.546	2.843	3.421	4.126	3.234

Table 19

Weighted Root Mean Squared Differences of Unrounded Scale Scores for Italian Using Traditional Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.751	0.832	2.025	1.870	1.370
FEE	0.661	0.717	1.836	1.739	1.238
2012-2014					
CHE	0.918	0.802	1.427	1.967	1.278
FEE	0.543	0.637	0.995	1.883	1.015
2013-2015					
CHE	0.931	1.160	1.653	1.678	1.355
FEE	0.739	0.892	1.317	1.620	1.142
Average	0.757	0.840	1.542	1.793	1.233
			ES = 0.1		
2012-2013					
CHE	0.867	0.957	1.884	1.898	1.401
FEE	0.975	1.000	1.715	1.853	1.386
2012-2014					
CHE	0.913	0.831	1.289	1.970	1.251
FEE	0.677	0.824	1.090	1.950	1.136
2013-2015					
CHE	1.065	1.255	1.412	1.692	1.356
FEE	1.142	1.294	1.239	1.627	1.326
Average	0.940	1.027	1.438	1.832	1.309
			ES = 0.3		
2012-2013					
CHE	1.106	1.113	1.999	1.956	1.543
FEE	1.414	1.360	1.877	1.981	1.658
2012-2014					
CHE	0.918	0.909	1.174	1.859	1.215
FEE	0.982	1.082	1.172	1.877	1.278
2013-2015					
CHE	1.355	1.657	1.548	1.772	1.583
FEE	2.239	2.318	1.714	1.850	2.030
Average	1.336	1.406	1.581	1.882	1.551

Table 20

Weighted Root Mean Squared Differences of Unrounded Composite Scores for French Using IRT Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.476	1.025	1.336	0.813	0.476
IRT TS	0.505	1.017	1.315	0.815	0.505
2012-2014					
IRT OS	1.220	0.880	1.050	1.021	1.220
IRT TS	1.259	0.883	1.034	1.030	1.259
2013-2015					
IRT OS	0.981	2.029	0.677	1.246	0.981
IRT TS	1.039	2.139	0.784	1.339	1.039
Average	0.913	1.329	1.033	1.044	0.913
			ES = 0.1		
2012-2013					
IRT OS	0.483	0.624	0.692	1.377	0.794
IRT TS	0.508	0.656	0.687	1.365	0.804
2012-2014					
IRT OS	1.173	1.422	0.911	1.096	1.151
IRT TS	1.193	1.446	0.942	1.077	1.164
2013-2015					
IRT OS	1.533	1.247	2.199	0.781	1.440
IRT TS	1.628	1.298	2.300	0.857	1.521
Average	1.086	1.116	1.289	1.092	1.146
			ES = 0.3		
2012-2013					
IRT OS	0.869	0.882	0.760	1.448	0.990
IRT TS	0.911	0.927	0.770	1.444	1.013
2012-2014					
IRT OS	1.886	1.498	0.881	0.591	1.214
IRT TS	1.876	1.510	0.898	0.587	1.218
2013-2015					
IRT OS	1.073	0.673	2.026	0.707	1.120
IRT TS	1.084	0.691	2.065	0.746	1.147
Average	1.283	1.030	1.234	0.921	1.117

Table 21

Weighted Root Mean Squared Differences of Unrounded Scale Scores for French Using IRT Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.243	0.205	0.464	0.581	0.373
IRT TS	0.261	0.235	0.471	0.579	0.387
2012-2014					
IRT OS	0.463	0.599	0.394	0.524	0.495
IRT TS	0.464	0.612	0.394	0.508	0.495
2013-2015					
IRT OS	0.704	0.446	1.095	0.357	0.651
IRT TS	0.747	0.470	1.136	0.408	0.690
Average	0.480	0.428	0.659	0.493	0.515
			ES = 0.1		
2012-2013					
IRT OS	0.220	0.279	0.336	0.611	0.362
IRT TS	0.252	0.308	0.351	0.615	0.382
2012-2014					
IRT OS	0.608	0.601	0.474	0.500	0.546
IRT TS	0.609	0.612	0.482	0.483	0.546
2013-2015					
IRT OS	0.833	0.550	1.195	0.384	0.740
IRT TS	0.871	0.577	1.227	0.430	0.776
Average	0.566	0.488	0.677	0.504	0.559
			ES = 0.3		
2012-2013					
IRT OS	0.417	0.439	0.406	0.672	0.484
IRT TS	0.464	0.484	0.447	0.693	0.522
2012-2014					
IRT OS	0.933	0.620	0.487	0.258	0.575
IRT TS	0.918	0.622	0.485	0.249	0.568
2013-2015					
IRT OS	0.550	0.303	1.084	0.372	0.577
IRT TS	0.556	0.313	1.090	0.389	0.587
Average	0.640	0.464	0.666	0.439	0.552

Table 22

Weighted Root Mean Squared Differences of Unrounded Composite Scores for French Using Traditional Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.661	0.572	2.223	2.423	1.470
FEE	0.480	0.462	2.007	2.357	1.327
2012-2014					
CHE	0.786	0.952	1.965	2.154	1.464
FEE	0.697	0.807	1.746	2.054	1.326
2013-2015					
CHE	1.379	1.160	2.213	2.371	1.781
FEE	1.222	0.995	1.889	2.272	1.595
Average	0.871	0.825	2.007	2.272	1.494
			ES = 0.1		
2012-2013					
CHE	0.838	0.802	2.118	2.509	1.567
FEE	0.643	0.608	1.624	2.363	1.310
2012-2014					
CHE	1.153	1.118	1.752	2.028	1.513
FEE	1.402	1.287	1.319	1.874	1.471
2013-2015					
CHE	1.497	1.263	1.978	2.465	1.801
FEE	1.646	0.936	1.628	2.301	1.628
Average	1.196	1.002	1.737	2.257	1.548
			ES = 0.3		
2012-2013					
CHE	0.960	0.815	1.877	2.522	1.544
FEE	1.607	1.150	1.063	2.260	1.520
2012-2014					
CHE	1.871	1.495	1.155	1.601	1.531
FEE	2.877	2.245	0.386	1.327	1.709
2013-2015					
CHE	1.991	0.806	1.665	2.175	1.659
FEE	3.799	2.046	1.775	1.830	2.362
Average	2.184	1.426	1.320	1.953	1.721

Table 23

Weighted Root Mean Squared Differences of Unrounded Scale Scores for French Using Traditional Equating Methods

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.419	0.371	1.112	1.166	0.767
FEE	0.336	0.331	0.964	1.123	0.688
2012-2014					
CHE	0.373	0.503	0.951	1.153	0.745
FEE	0.301	0.349	0.812	1.043	0.626
2013-2015					
CHE	0.657	0.628	0.994	1.059	0.835
FEE	0.628	0.563	0.847	1.020	0.764
Average	0.452	0.457	0.947	1.094	0.738
			ES = 0.1		
2012-2013					
CHE	0.512	0.529	1.061	1.198	0.825
FEE	0.275	0.289	0.774	1.097	0.609
2012-2014					
CHE	0.511	0.529	0.880	1.039	0.740
FEE	0.602	0.528	0.611	0.929	0.668
2013-2015					
CHE	0.696	0.711	0.881	1.093	0.845
FEE	0.733	0.431	0.755	1.025	0.736
Average	0.555	0.503	0.827	1.063	0.737
			ES = 0.3		
2012-2013					
CHE	0.449	0.460	0.919	1.199	0.757
FEE	0.653	0.472	0.520	1.056	0.675
2012-2014					
CHE	0.833	0.614	0.567	0.782	0.699
FEE	1.265	0.953	0.215	0.642	0.769
2013-2015					
CHE	0.911	0.352	0.834	0.931	0.757
FEE	1.680	0.905	0.953	0.793	1.083
Average	0.965	0.626	0.668	0.901	0.790

Table 24

Averages of Weighted Root Mean Squared Differences Across Equating Methods and Years

Subject	Score Type	Effect Size	MC-only	MC + Best FR	MC + Worst FR	MC + All FR
German	Unrounded composite raw score	0	1.815	2.390	3.022	3.522
		0.1	1.600	2.193	2.761	3.461
		0.3	2.668	2.856	3.784	3.818
German	Unrounded scale score	0	0.786	1.008	1.155	1.342
		0.1	0.685	0.919	1.057	1.315
		0.3	1.078	1.150	1.436	1.442
Italian	Unrounded composite raw score	0	1.496	2.067	3.298	4.079
		0.1	1.614	2.226	2.958	4.248
		0.3	2.287	2.990	3.207	4.163
Italian	Unrounded scale score	0	0.738	0.952	1.402	1.787
		0.1	0.818	1.038	1.282	1.847
		0.3	1.105	1.352	1.390	1.828
French	Unrounded composite raw score	0	0.892	1.077	1.520	1.658
		0.1	1.141	1.059	1.513	1.675
		0.3	1.734	1.228	1.277	1.437
French	Unrounded scale score	0	0.466	0.443	0.803	0.794
		0.1	0.561	0.496	0.752	0.784
		0.3	0.803	0.545	0.667	0.670

Table 25

Proportion of Equated Composite Scores within the DTM Boundaries for German for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.103	0.103	0.119	0.095	0.105
IRT TS	0.183	0.198	0.190	0.183	0.188
2012-2014					
IRT OS	0.167	0.103	0.175	0.103	0.137
IRT TS	0.254	0.190	0.262	0.206	0.228
2013-2015					
IRT OS	0.397	0.230	0.198	0.849	0.419
IRT TS	0.341	0.317	0.286	0.873	0.454
Average	0.241	0.190	0.205	0.385	0.255
			ES = 0.1		
2012-2013					
IRT OS	0.095	0.087	0.111	0.087	0.095
IRT TS	0.214	0.198	0.214	0.190	0.204
2012-2014					
IRT OS	0.230	0.095	0.167	0.111	0.151
IRT TS	0.262	0.198	0.246	0.222	0.232
2013-2015					
IRT OS	0.611	0.841	0.310	1.000	0.690
IRT TS	0.595	0.833	0.389	1.000	0.704
Average	0.335	0.376	0.239	0.435	0.346
			ES = 0.3		
2012-2013					
IRT OS	0.087	0.135	0.079	0.079	0.095
IRT TS	0.190	0.246	0.183	0.183	0.200
2012-2014					
IRT OS	0.103	0.079	0.103	0.111	0.099
IRT TS	0.159	0.143	0.151	0.159	0.153
2013-2015					
IRT OS	0.333	0.516	0.167	0.548	0.391
IRT TS	0.333	0.516	0.238	0.595	0.421
Average	0.201	0.272	0.153	0.279	0.227

Table 26

Proportion of Equated Scale Scores within the DTM Boundaries for German for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.246	0.278	0.262	0.238	0.256
IRT TS	0.270	0.302	0.270	0.262	0.276
2012-2014					
IRT OS	0.365	0.222	0.278	0.151	0.254
IRT TS	0.405	0.270	0.325	0.198	0.300
2013-2015					
IRT OS	0.746	0.865	0.452	0.968	0.758
IRT TS	0.754	0.881	0.500	0.984	0.780
Average	0.464	0.470	0.348	0.467	0.437
			ES = 0.1		
2012-2013					
IRT OS	0.262	0.270	0.302	0.238	0.268
IRT TS	0.286	0.286	0.302	0.270	0.286
2012-2014					
IRT OS	0.444	0.238	0.302	0.159	0.286
IRT TS	0.460	0.262	0.349	0.190	0.315
2013-2015					
IRT OS	0.944	0.937	0.690	1.000	0.893
IRT TS	0.944	0.952	0.714	1.000	0.903
Average	0.557	0.491	0.443	0.476	0.492
			ES = 0.3		
2012-2013					
IRT OS	0.254	0.341	0.222	0.206	0.256
IRT TS	0.278	0.397	0.230	0.238	0.286
2012-2014					
IRT OS	0.278	0.190	0.246	0.159	0.218
IRT TS	0.333	0.206	0.238	0.167	0.236
2013-2015					
IRT OS	0.810	0.873	0.333	0.937	0.738
IRT TS	0.794	0.841	0.357	0.952	0.736
Average	0.458	0.475	0.271	0.443	0.412

Table 27

Proportion of Equated Composite Scores within the DTM Boundaries for German for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.143	0.087	0.087	0.135	0.113
FEE	0.206	0.111	0.119	0.151	0.147
2012-2014					
CHE	0.159	0.119	0.056	0.048	0.095
FEE	0.238	0.167	0.040	0.063	0.127
2013-2015					
CHE	0.151	0.119	0.063	0.381	0.179
FEE	0.230	0.087	0.159	0.294	0.192
Average	0.188	0.115	0.087	0.179	0.142
			ES = 0.1		
2012-2013					
CHE	0.087	0.095	0.071	0.119	0.093
FEE	0.262	0.095	0.119	0.143	0.155
2012-2014					
CHE	0.198	0.135	0.063	0.048	0.111
FEE	0.421	0.222	0.048	0.056	0.187
2013-2015					
CHE	0.198	0.214	0.063	0.175	0.163
FEE	0.325	0.222	0.056	0.103	0.177
Average	0.249	0.164	0.070	0.107	0.147
			ES = 0.3		
2012-2013					
CHE	0.079	0.056	0.071	0.183	0.097
FEE	0.167	0.127	0.111	0.19	0.149
2012-2014					
CHE	0.167	0.111	0.063	0.04	0.095
FEE	0.206	0.175	0.048	0.04	0.117
2013-2015					
CHE	0.222	0.183	0.143	0.103	0.163
FEE	0.079	0.317	0.143	0.079	0.155
Average	0.153	0.161	0.097	0.106	0.129

Table 28

Proportion of Equated Scale Scores within the DTM Boundaries for German for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.405	0.357	0.317	0.365	0.361
FEE	0.476	0.365	0.278	0.365	0.371
2012-2014					
CHE	0.444	0.286	0.135	0.111	0.244
FEE	0.532	0.349	0.103	0.103	0.272
2013-2015					
CHE	0.413	0.246	0.175	0.429	0.315
FEE	0.500	0.365	0.206	0.429	0.375
Average	0.462	0.328	0.202	0.300	0.323
			ES = 0.1		
2012-2013					
CHE	0.365	0.349	0.246	0.373	0.333
FEE	0.444	0.333	0.254	0.381	0.353
2012-2014					
CHE	0.516	0.302	0.151	0.095	0.266
FEE	0.786	0.381	0.119	0.063	0.337
2013-2015					
CHE	0.476	0.333	0.183	0.452	0.361
FEE	0.540	0.548	0.143	0.437	0.417
Average	0.521	0.374	0.183	0.300	0.345
			ES = 0.3		
2012-2013					
CHE	0.365	0.333	0.365	0.381	0.361
FEE	0.413	0.421	0.222	0.373	0.357
2012-2014					
CHE	0.381	0.325	0.087	0.095	0.222
FEE	0.349	0.302	0.087	0.087	0.206
2013-2015					
CHE	0.460	0.389	0.413	0.429	0.423
FEE	0.230	0.444	0.135	0.429	0.310
Average	0.366	0.369	0.218	0.299	0.313

Table 29

Proportion of Equated Composite Scores within the DTM Boundaries for Italian for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.359	0.305	0.092	0.107	0.216
IRT TS	0.450	0.397	0.191	0.206	0.311
2012-2014					
IRT OS	0.290	0.244	0.244	0.168	0.237
IRT TS	0.214	0.176	0.183	0.145	0.179
2013-2015					
IRT OS	0.496	0.229	0.214	0.611	0.387
IRT TS	0.511	0.260	0.282	0.641	0.424
Average	0.387	0.268	0.201	0.313	0.292
			ES = 0.1		
2012-2013					
IRT OS	0.214	0.198	0.145	0.206	0.191
IRT TS	0.290	0.229	0.214	0.237	0.242
2012-2014					
IRT OS	0.359	0.290	0.260	0.145	0.263
IRT TS	0.328	0.252	0.244	0.176	0.250
2013-2015					
IRT OS	0.557	0.206	0.191	0.336	0.323
IRT TS	0.534	0.160	0.183	0.374	0.313
Average	0.380	0.223	0.206	0.246	0.264
			ES = 0.3		
2012-2013					
IRT OS	0.511	0.290	0.107	0.176	0.271
IRT TS	0.557	0.321	0.198	0.206	0.321
2012-2014					
IRT OS	0.366	0.137	0.313	0.229	0.261
IRT TS	0.389	0.191	0.336	0.229	0.286
2013-2015					
IRT OS	0.115	0.099	0.153	0.290	0.164
IRT TS	0.122	0.107	0.122	0.321	0.168
Average	0.344	0.191	0.205	0.242	0.245

Table 30

Proportion of Equated Scale Scores within the DTM Boundaries for Italian for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.748	0.695	0.267	0.267	0.494
IRT TS	0.740	0.695	0.305	0.313	0.513
2012-2014					
IRT OS	0.427	0.328	0.328	0.229	0.328
IRT TS	0.412	0.305	0.305	0.237	0.315
2013-2015					
IRT OS	0.794	0.481	0.290	0.931	0.624
IRT TS	0.786	0.511	0.344	0.969	0.653
Average	0.651	0.503	0.307	0.491	0.488
			ES = 0.1		
2012-2013					
IRT OS	0.763	0.756	0.321	0.290	0.532
IRT TS	0.802	0.748	0.359	0.298	0.552
2012-2014					
IRT OS	0.450	0.336	0.374	0.252	0.353
IRT TS	0.435	0.359	0.382	0.237	0.353
2013-2015					
IRT OS	0.771	0.405	0.267	0.863	0.576
IRT TS	0.802	0.382	0.275	0.908	0.592
Average	0.670	0.497	0.330	0.475	0.493
			ES = 0.3		
2012-2013					
IRT OS	0.725	0.626	0.366	0.214	0.483
IRT TS	0.763	0.664	0.397	0.267	0.523
2012-2014					
IRT OS	0.458	0.214	0.382	0.290	0.336
IRT TS	0.450	0.282	0.382	0.305	0.355
2013-2015					
IRT OS	0.443	0.198	0.145	0.748	0.384
IRT TS	0.450	0.198	0.160	0.779	0.397
Average	0.548	0.364	0.305	0.434	0.413

Table 31

Proportion of Equated Composite Scores within the DTM Boundaries for Italian for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.244	0.290	0.038	0.084	0.164
FEE	0.176	0.351	0.038	0.084	0.162
2012-2014					
CHE	0.282	0.321	0.038	0.038	0.170
FEE	0.313	0.359	0.061	0.053	0.197
2013-2015					
CHE	0.206	0.206	0.038	0.176	0.156
FEE	0.221	0.252	0.046	0.168	0.172
Average	0.240	0.296	0.043	0.101	0.170
			ES = 0.1		
2012-2013					
CHE	0.214	0.206	0.130	0.084	0.158
FEE	0.344	0.305	0.115	0.099	0.216
2012-2014					
CHE	0.176	0.198	0.092	0.031	0.124
FEE	0.458	0.313	0.053	0.038	0.216
2013-2015					
CHE	0.168	0.145	0.176	0.145	0.158
FEE	0.229	0.000	0.084	0.160	0.118
Average	0.265	0.195	0.108	0.093	0.165
			ES = 0.3		
2012-2013					
CHE	0.145	0.214	0.137	0.115	0.153
FEE	0.351	0.298	0.183	0.145	0.244
2012-2014					
CHE	0.122	0.229	0.237	0.038	0.156
FEE	0.191	0.191	0.183	0.046	0.153
2013-2015					
CHE	0.267	0.031	0.099	0.122	0.130
FEE	0.053	0.122	0.221	0.137	0.134
Average	0.188	0.181	0.177	0.101	0.162

Table 32

Proportion of Equated Scale Scores within the DTM Boundaries for Italian for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.595	0.618	0.221	0.282	0.429
FEE	0.664	0.679	0.214	0.336	0.473
2012-2014					
CHE	0.687	0.618	0.366	0.168	0.460
FEE	0.718	0.672	0.427	0.176	0.498
2013-2015					
CHE	0.557	0.573	0.237	0.534	0.475
FEE	0.664	0.611	0.260	0.542	0.519
Average	0.648	0.628	0.288	0.340	0.476
			ES = 0.1		
2012-2013					
CHE	0.679	0.641	0.282	0.336	0.485
FEE	0.611	0.588	0.290	0.344	0.458
2012-2014					
CHE	0.595	0.580	0.420	0.183	0.445
FEE	0.687	0.588	0.443	0.160	0.469
2013-2015					
CHE	0.511	0.382	0.305	0.450	0.412
FEE	0.435	0.344	0.313	0.458	0.387
Average	0.587	0.520	0.342	0.322	0.443
			ES = 0.3		
2012-2013					
CHE	0.595	0.557	0.321	0.374	0.462
FEE	0.550	0.519	0.328	0.374	0.443
2012-2014					
CHE	0.405	0.420	0.435	0.214	0.368
FEE	0.496	0.389	0.397	0.191	0.368
2013-2015					
CHE	0.489	0.252	0.305	0.504	0.387
FEE	0.336	0.115	0.359	0.489	0.324
Average	0.478	0.375	0.358	0.358	0.392

Table 33

Proportion of Equated Composite Scores within the DTM Boundaries for French for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.706	0.754	0.302	0.198	0.490
IRT TS	0.738	0.722	0.381	0.230	0.518
2012-2014					
IRT OS	0.214	0.222	0.437	0.056	0.232
IRT TS	0.238	0.198	0.452	0.071	0.240
2013-2015					
IRT OS	0.159	0.278	0.127	0.397	0.240
IRT TS	0.230	0.365	0.190	0.405	0.298
Average	0.381	0.423	0.315	0.226	0.336
			ES = 0.1		
2012-2013					
IRT OS	0.698	0.603	0.254	0.127	0.421
IRT TS	0.627	0.516	0.262	0.175	0.395
2012-2014					
IRT OS	0.151	0.317	0.270	0.127	0.216
IRT TS	0.183	0.270	0.278	0.175	0.226
2013-2015					
IRT OS	0.167	0.254	0.127	0.270	0.204
IRT TS	0.230	0.325	0.183	0.286	0.256
Average	0.343	0.381	0.229	0.193	0.286
			ES = 0.3		
2012-2013					
IRT OS	0.397	0.381	0.286	0.103	0.292
IRT TS	0.317	0.31	0.333	0.143	0.276
2012-2014					
IRT OS	0.357	0.302	0.151	0.413	0.306
IRT TS	0.389	0.381	0.286	0.548	0.401
2013-2015					
IRT OS	0.286	0.437	0.159	0.357	0.31
IRT TS	0.302	0.437	0.214	0.373	0.331
Average	0.341	0.374	0.238	0.323	0.319

Table 34

Proportion of Equated Scale Scores within the DTM Boundaries for French for IRT Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.849	0.889	0.444	0.333	0.629
IRT TS	0.825	0.857	0.476	0.405	0.641
2012-2014					
IRT OS	0.437	0.294	0.524	0.413	0.417
IRT TS	0.437	0.310	0.532	0.421	0.425
2013-2015					
IRT OS	0.294	0.524	0.198	0.603	0.405
IRT TS	0.317	0.540	0.230	0.540	0.407
Average	0.526	0.569	0.401	0.452	0.487
			ES = 0.1		
2012-2013					
IRT OS	0.873	0.889	0.595	0.294	0.663
IRT TS	0.825	0.833	0.595	0.286	0.635
2012-2014					
IRT OS	0.389	0.405	0.397	0.429	0.405
IRT TS	0.397	0.405	0.405	0.476	0.421
2013-2015					
IRT OS	0.270	0.333	0.206	0.571	0.345
IRT TS	0.278	0.357	0.214	0.484	0.333
Average	0.505	0.537	0.402	0.423	0.467
			ES = 0.3		
2012-2013					
IRT OS	0.611	0.595	0.587	0.063	0.464
IRT TS	0.548	0.524	0.635	0.079	0.446
2012-2014					
IRT OS	0.421	0.317	0.421	0.714	0.468
IRT TS	0.429	0.381	0.444	0.810	0.516
2013-2015					
IRT OS	0.437	0.659	0.238	0.556	0.472
IRT TS	0.460	0.675	0.294	0.571	0.500
Average	0.484	0.525	0.437	0.466	0.478

Table 35

Proportion of Equated Composite Scores within the DTM Boundaries for French for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.381	0.683	0.063	0.079	0.302
FEE	0.516	0.619	0.056	0.063	0.313
2012-2014					
CHE	0.325	0.357	0.095	0.024	0.200
FEE	0.587	0.571	0.063	0.024	0.312
2013-2015					
CHE	0.175	0.373	0.175	0.270	0.248
FEE	0.190	0.294	0.190	0.286	0.240
Average	0.362	0.483	0.107	0.124	0.269
			ES = 0.1		
2012-2013					
CHE	0.302	0.437	0.087	0.048	0.218
FEE	0.579	0.468	0.095	0.063	0.302
2012-2014					
CHE	0.294	0.413	0.016	0.016	0.185
FEE	0.413	0.333	0.127	0.024	0.224
2013-2015					
CHE	0.175	0.214	0.310	0.183	0.220
FEE	0.206	0.357	0.206	0.198	0.242
Average	0.328	0.370	0.140	0.089	0.232
			ES = 0.3		
2012-2013					
CHE	0.198	0.246	0.032	0.024	0.125
FEE	0.302	0.381	0.079	0.040	0.200
2012-2014					
CHE	0.111	0.310	0.056	0.048	0.131
FEE	0.056	0.063	0.532	0.071	0.181
2013-2015					
CHE	0.056	0.492	0.190	0.286	0.256
FEE	0.143	0.175	0.103	0.476	0.224
Average	0.144	0.278	0.165	0.157	0.186

Table 36

Proportion of Equated Scale Scores within the DTM Boundaries for French for Traditional Equating Conditions

		Composition of Common Item Set			
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.762	0.825	0.127	0.111	0.456
FEE	0.810	0.810	0.159	0.151	0.482
2012-2014					
CHE	0.762	0.667	0.183	0.087	0.425
FEE	0.952	0.937	0.222	0.119	0.558
2013-2015					
CHE	0.397	0.484	0.341	0.373	0.399
FEE	0.405	0.524	0.333	0.405	0.417
Average	0.681	0.708	0.228	0.208	0.456
			ES = 0.1		
2012-2013					
CHE	0.714	0.746	0.183	0.135	0.444
FEE	0.944	0.889	0.230	0.151	0.554
2012-2014					
CHE	0.587	0.667	0.278	0.143	0.419
FEE	0.635	0.675	0.357	0.183	0.462
2013-2015					
CHE	0.397	0.484	0.579	0.373	0.458
FEE	0.357	0.683	0.492	0.389	0.480
Average	0.606	0.690	0.353	0.229	0.470
			ES = 0.3		
2012-2013					
CHE	0.571	0.619	0.063	0.079	0.333
FEE	0.524	0.643	0.460	0.087	0.429
2012-2014					
CHE	0.310	0.444	0.389	0.310	0.363
FEE	0.159	0.167	0.802	0.381	0.377
2013-2015					
CHE	0.183	0.762	0.357	0.540	0.460
FEE	0.254	0.294	0.222	0.556	0.331
Average	0.333	0.488	0.382	0.325	0.382

Table 37

Proportion of Equated Scores within DTM Boundary Averaged Across Equating Methods and Years

Subject	Score Type	Effect Size	MC-only	MC + Best FR	MC + Worst FR	MC + All FR
German	Unrounded composite raw score	0	0.215	0.153	0.146	0.282
		0.1	0.292	0.270	0.155	0.271
		0.3	0.177	0.217	0.125	0.193
German	Unrounded scale score	0	0.463	0.399	0.275	0.384
		0.1	0.539	0.433	0.313	0.388
		0.3	0.412	0.422	0.245	0.371
Italian	Unrounded composite raw score	0	0.314	0.282	0.122	0.207
		0.1	0.323	0.209	0.157	0.170
		0.3	0.266	0.186	0.191	0.172
Italian	Unrounded scale score	0	0.650	0.566	0.298	0.416
		0.1	0.629	0.509	0.336	0.399
		0.3	0.513	0.370	0.332	0.396
French	Unrounded composite raw score	0	0.372	0.453	0.211	0.175
		0.1	0.336	0.376	0.185	0.141
		0.3	0.243	0.326	0.202	0.240
French	Unrounded scale score	0	0.604	0.639	0.315	0.330
		0.1	0.556	0.614	0.378	0.326
		0.3	0.409	0.507	0.410	0.396

Table 38

Classification Consistency for German Using IRT Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.919	0.906	0.879	0.885	0.897
IRT TS	0.920	0.918	0.880	0.897	0.904
2012-2014					
IRT OS	0.941	0.863	0.887	0.793	0.871
IRT TS	0.948	0.850	0.876	0.793	0.867
2013-2015					
IRT OS	0.971	0.984	0.969	1.000	0.981
IRT TS	0.971	0.984	0.969	0.994	0.980
Average	0.945	0.918	0.910	0.894	0.917
			ES = 0.1		
2012-2013					
IRT OS	0.927	0.896	0.903	0.892	0.905
IRT TS	0.939	0.908	0.915	0.910	0.918
2012-2014					
IRT OS	0.949	0.871	0.924	0.798	0.886
IRT TS	0.949	0.871	0.924	0.798	0.886
2013-2015					
IRT OS	0.975	1.000	0.994	0.977	0.987
IRT TS	0.975	1.000	0.980	0.986	0.985
Average	0.952	0.924	0.940	0.894	0.928
			ES = 0.3		
2012-2013					
IRT OS	0.917	0.929	0.875	0.892	0.903
IRT TS	0.929	0.941	0.868	0.904	0.911
2012-2014					
IRT OS	0.930	0.839	0.877	0.764	0.853
IRT TS	0.924	0.834	0.871	0.764	0.848
2013-2015					
IRT OS	0.973	0.984	0.939	0.986	0.971
IRT TS	0.963	0.984	0.929	0.986	0.966
Average	0.939	0.919	0.893	0.883	0.908

Table 39

Classification Consistency for German using Traditional Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.938	0.914	0.878	0.897	0.907
FEE	0.956	0.945	0.895	0.909	0.926
2012-2014					
CHE	0.941	0.903	0.854	0.781	0.870
FEE	0.956	0.932	0.847	0.792	0.882
2013-2015					
CHE	0.895	0.885	0.827	0.877	0.871
FEE	0.934	0.906	0.843	0.893	0.894
Average	0.937	0.914	0.857	0.858	0.892
			ES = 0.1		
2012-2013					
CHE	0.945	0.914	0.874	0.897	0.908
FEE	0.956	0.931	0.874	0.897	0.915
2012-2014					
CHE	0.928	0.913	0.854	0.792	0.872
FEE	0.953	0.932	0.847	0.792	0.881
2013-2015					
CHE	0.902	0.895	0.832	0.859	0.872
FEE	0.931	0.923	0.827	0.859	0.885
Average	0.936	0.918	0.851	0.849	0.889
			ES = 0.3		
2012-2013					
CHE	0.925	0.919	0.859	0.897	0.900
FEE	0.887	0.934	0.849	0.897	0.892
2012-2014					
CHE	0.919	0.893	0.805	0.777	0.849
FEE	0.908	0.864	0.792	0.777	0.835
2013-2015					
CHE	0.875	0.872	0.817	0.854	0.855
FEE	0.832	0.893	0.783	0.854	0.841
Average	0.891	0.896	0.818	0.843	0.862

Table 40

Classification Consistency for Italian Using IRT Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
IRT OS	0.991	0.962	0.918	0.836	0.927
IRT TS	0.991	0.971	0.918	0.836	0.929
2012-2014					
IRT OS	0.890	0.838	0.838	0.716	0.821
IRT TS	0.890	0.838	0.838	0.706	0.818
2013-2015					
IRT OS	0.958	0.913	0.892	0.985	0.937
IRT TS	0.958	0.913	0.892	0.996	0.940
Average	0.946	0.906	0.883	0.846	0.895
			ES = 0.1		
2012-2013					
IRT OS	0.962	0.952	0.935	0.809	0.915
IRT TS	0.952	0.952	0.935	0.809	0.912
2012-2014					
IRT OS	0.887	0.842	0.856	0.716	0.825
IRT TS	0.887	0.842	0.856	0.721	0.827
2013-2015					
IRT OS	0.969	0.914	0.903	0.974	0.940
IRT TS	0.969	0.914	0.903	0.974	0.940
Average	0.938	0.903	0.898	0.834	0.893
			ES = 0.3		
2012-2013					
IRT OS	0.969	0.955	0.924	0.821	0.917
IRT TS	0.969	0.945	0.924	0.821	0.915
2012-2014					
IRT OS	0.895	0.807	0.850	0.716	0.817
IRT TS	0.895	0.807	0.850	0.716	0.817
2013-2015					
IRT OS	0.925	0.845	0.870	0.981	0.905
IRT TS	0.910	0.865	0.860	0.992	0.907
Average	0.927	0.871	0.880	0.841	0.880

Table 41

Classification Consistency for Italian using Traditional Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.972	0.974	0.821	0.863	0.908
FEE	0.967	0.962	0.836	0.863	0.907
2012-2014					
CHE	0.969	0.967	0.921	0.830	0.922
FEE	0.969	0.983	0.932	0.829	0.928
2013-2015					
CHE	0.950	0.958	0.852	0.914	0.919
FEE	0.970	0.946	0.884	0.900	0.925
Average	0.966	0.965	0.874	0.867	0.918
			ES = 0.1		
2012-2013					
CHE	0.970	0.969	0.825	0.866	0.908
FEE	0.969	0.969	0.852	0.866	0.914
2012-2014					
CHE	0.941	0.938	0.932	0.830	0.910
FEE	0.954	0.946	0.932	0.830	0.916
2013-2015					
CHE	0.954	0.93	0.881	0.914	0.920
FEE	0.940	0.904	0.902	0.914	0.915
Average	0.955	0.943	0.887	0.870	0.914
			ES = 0.3		
2012-2013					
CHE	0.949	0.969	0.825	0.866	0.902
FEE	0.950	0.940	0.835	0.866	0.898
2012-2014					
CHE	0.933	0.914	0.925	0.830	0.901
FEE	0.930	0.921	0.921	0.830	0.901
2013-2015					
CHE	0.940	0.881	0.877	0.902	0.900
FEE	0.871	0.838	0.897	0.902	0.877
Average	0.929	0.911	0.880	0.866	0.896

Table 42

Classification Consistency for French Using IRT Equating Methods

Years/Method	Composition of Common Item Set: MC-only	MC + Best FR	MC + Worst FR	MC + All FR	Average
		ES = 0			
2012-2013					
IRT OS	0.952	0.987	0.976	0.956	0.968
IRT TS	0.969	0.982	0.980	0.961	0.973
2012-2014					
IRT OS	0.972	0.962	0.978	0.954	0.967
IRT TS	0.967	0.962	0.978	0.954	0.965
2013-2015					
IRT OS	0.915	0.973	0.888	0.982	0.940
IRT TS	0.915	0.973	0.888	0.982	0.940
Average	0.948	0.973	0.948	0.965	0.959
		ES = 0.1			
2012-2013					
IRT OS	0.987	0.987	0.991	0.941	0.977
IRT TS	0.982	0.982	0.996	0.961	0.980
2012-2014					
IRT OS	0.957	0.924	0.967	0.954	0.951
IRT TS	0.945	0.942	0.967	0.954	0.952
2013-2015					
IRT OS	0.924	0.959	0.902	0.982	0.942
IRT TS	0.916	0.947	0.892	0.971	0.932
Average	0.952	0.957	0.953	0.961	0.955
		ES = 0.3			
2012-2013					
IRT OS	0.967	0.967	0.961	0.946	0.960
IRT TS	0.962	0.962	0.966	0.950	0.960
2012-2014					
IRT OS	0.923	0.941	0.967	0.982	0.953
IRT TS	0.928	0.941	0.954	0.982	0.951
2013-2015					
IRT OS	0.932	0.994	0.893	0.968	0.947
IRT TS	0.932	0.982	0.893	0.956	0.941
Average	0.941	0.965	0.939	0.964	0.952

Table 43

Classification Consistency for French using Traditional Equating Methods

	Composition of Common Item Set				
Years/Method	**MC-only**	**MC + Best FR**	**MC + Worst FR**	**MC + All FR**	**Average**
			ES = 0		
2012-2013					
CHE	0.981	0.985	0.894	0.882	0.936
FEE	0.996	0.996	0.913	0.894	0.950
2012-2014					
CHE	0.965	0.957	0.921	0.888	0.933
FEE	0.965	0.957	0.921	0.921	0.941
2013-2015					
CHE	0.933	0.953	0.910	0.926	0.931
FEE	0.940	0.949	0.914	0.942	0.936
Average	0.963	0.966	0.912	0.909	0.938
			ES = 0.1		
2012-2013					
CHE	0.968	0.972	0.898	0.894	0.933
FEE	0.987	0.987	0.924	0.894	0.948
2012-2014					
CHE	0.938	0.970	0.933	0.916	0.939
FEE	0.943	0.938	0.961	0.921	0.941
2013-2015					
CHE	0.925	0.969	0.933	0.926	0.938
FEE	0.914	0.973	0.929	0.931	0.937
Average	0.946	0.968	0.930	0.914	0.939
			ES = 0.3		
2012-2013					
CHE	0.962	0.968	0.902	0.886	0.930
FEE	0.933	0.963	0.950	0.898	0.936
2012-2014					
CHE	0.910	0.949	0.966	0.933	0.940
FEE	0.874	0.901	0.995	0.945	0.929
2013-2015					
CHE	0.921	0.970	0.929	0.942	0.941
FEE	0.815	0.913	0.913	0.948	0.897
Average	0.903	0.944	0.943	0.925	0.929

Table 44

Classification Consistency Indices Averaged Over Equating Methods and Years

Subject	Effect Size	MC-only	MC + Best FR	MC + Worst FR	MC + All FR
German	0	0.941	0.916	0.884	0.876
	0.1	0.944	0.921	0.896	0.872
	0.3	0.915	0.908	0.856	0.863
Italian	0	0.956	0.936	0.879	0.857
	0.1	0.947	0.923	0.893	0.852
	0.3	0.928	0.891	0.880	0.854
French	0	0.956	0.970	0.930	0.937
	0.1	0.949	0.963	0.942	0.938
	0.3	0.922	0.955	0.941	0.945

Chapter 4: Minimum Sample Size Needed for Equipercentile Equating under the Random Groups Design

Shichao Wang and Huan Liu
The University of Iowa, Iowa City, IA

Abstract

The main purpose of this study is to investigate the minimum sample size needed for equipercentile equating to obtain accurate equating results in the random groups (RG) design. Data were simulated based on the Advanced Placement (AP) US History Examinations from College Board, which are mixed-format tests containing a multiple-choice (MC) section and a free-response (FR) section. Three-parameter logistic and graded response models were used to generate data for MC and FR section, respectively. This study investigated the impacts of three factors on the accuracy of equipercentile equating results including a wide range of sample sizes, presmoothing and postsmoothing methods, and four levels of group ability difference when the randomly equivalent assumption is met (group difference of 0) or violated to some extent (group difference greater than 0). The criterion equating was defined as a single group equipercentile equating resulting from the generating item parameters and two large groups both simulated from the standard normal distribution. The results were consistent with previous studies that larger sample sizes produced more accurate equating results. The rate of increase in equating accuracy was more obvious in sample sizes ranging from 500 to 3,000, and less obvious for sample sizes from 3,000 to 8,000. The Difference That Matters (DTM) criterion suggested that sample sizes greater than or equal 3,000 produced acceptable amount of equating error when the assumptions of random groups design were met and the group difference was zero. If the randomly equivalent groups assumption is violated and group ability difference was larger than 0.05, the equating results would be bias no matter how large the sample size were used. Smoothing can improve equating accuracy. Postsmoothing led to slightly more accurate equating results than presmoothing.

Minimum Sample Size Needed for Equipercentile Equating under the Random Groups Design

According to Kolen and Brennan (2014), the definition of equating is a statistical process used for adjusting scores obtained from test forms so that those scores can be used interchangeably. Equating is an important process by which testing programs can maintain a score standard across various test forms (Skaggs, 2005).

The equating design considered in this study is the RG design. In this design, examinees are randomly and independently assigned to one form of the test. Compared to the single group design, in which examinees take more than one form, the RG design enables time to be minimized and thus can be applied more practically and efficiently. In the RG design, multiple forms are administered simultaneously, which lowers the item exposure rate and helps minimize test security concerns. Through a carefully implemented random assignment procedure, the analyses of the test forms and examinee performance are straightforward, as the differences in group performance on the different forms is a direct indication of the differences in difficulty between the forms (Kolen & Brennan, 2014).

Under the RG design, equipercentile equating is a widely used equating method. In equipercentile equating, examinees' scores on different forms are equated with respect to their corresponding percentile ranks. The detailed procedures of conducting equipercentile equating can be found in Kolen and Brennan (2014).

The main source of equating error comes from sampling error. Previous studies (Asiret & Sünbül, 2016; Cui & Kolen, 2009; Liu & Kolen, 2011b) showed that sampling error can be minimized by increasing sample size. Typically, the larger the sample size is, the better the sample represents the population, and thus the greater the accuracy of equating. However, it could be time consuming and costly for a testing organization to collect a large sample of data. Thus, it is challenging to determine the appropriate sample size to obtain acceptable equating results.

In addition to sample size, smoothing also affects the accuracy of equating. Presmoothing and postsmoothing are two approaches to smooth the estimated equating relationships obtained from sample statistics. In presmoothing, the observed score distributions are smoothed prior to conducting equating. The most commonly used presmoothing method is polynomial loglinear presmoothing. In postsmoothing, the equated scores are smoothed. Cubic spline postsmoothing is

the most widely used postsmoothing method. Both polynomial loglinear presmoothing and cubic spline postsmoothing have shown the potential to improve equating accuracy and provide comparably precise equating results (Liu & Kolen, 2011a).

The accuracy of equating results depends on how well the assumptions for equipercentile equating under the RG design are met as well. One important assumption under the random group design is that the ability of the group taking one form is randomly equivalent to the ability of the group taking the other form. In practice, the randomly equivalent groups are typically obtained through a spiraling process by distributing the test booklet alternately. If the spiraling process is not followed properly by the room supervisor, the assumption of randomly equivalent groups is violated to some extent, and thus there would be difference in the abilities of the two groups. When investigating the minimum sample size needed for equipercentile equating, the possibility of existence of group ability difference should be considered to assure the accuracy of equating.

The confounding effects of smoothing and group ability difference on the sample size needed for equipercentile equating to achieve adequate equating in the RG design are not fully understood. Thus, this study offers an intensively investigation of the impact of different levels of sample sizes on the accuracy of equipercentile equating method under various conditions for the RG design.

Method

Data

Real data from the AP US History Examinations were calibrated to obtain the generating item parameters for this study. One administration of AP US History includes two parallel mixed-format forms, and each form contains 49 MC and 5 FR items. The MC items were scored dichotomously and the FR items were each scored by human raters on a 6-point scale (i.e., 0, 1, 2, 3, 4, and 5). In this study, the MC and FR items were calibrated using flexMIRT 3.0 (Cai, 2015) based on the three-parameter logistic (3PL; Birnbaum, 1968) and graded response models, respectively. The flexMIRT default calibration settings were adopted with one exception: a prior distribution (i.e., $c \sim$ beta (5, 17)) was used to estimate the pseudo-guessing parameter. The estimated item parameters for the old and new forms were used to generate item responses based on the same IRT models: 3PL model for MC items and graded response model for FR items.

Factors of Investigation

Three factors of investigation were considered in this research: sample size, smoothing method, and group ability difference. All analyses were conducted using *Equating Recipes* (Brennan et al., 2009).

Sample size. The sensitivity of sample size on equating results is the focus of this study. To cover a wide range of possible sample sizes, 16 levels of sample sizes were examined, including sample sizes from 500 to 8,000 with an increment of 500.

Smoothing method. Two smoothing methods, loglinear presmoothing and cubic spline postsmoothing were considered in this study. In addition, equipercentile equating without smoothing was included as well. For loglinear presmoothing, the smoothing parameter was selected using the AIC for the old and new form groups. The parameter used for cubic spline postsmoothing was 0.1 to obtain a moderate smoothing effect.

Group ability difference. Group ability difference is examined to investigate the accuracy of equating results when the randomly equivalent groups assumption is violated. Pseudo-groups were used to achieve various levels of group ability differences. The pseudo groups were created by simulating examinees from different population distributions. Group ability difference was characterized by differences in mean of the generated population distribution. The old form was assumed to be taken by a typical examinee group following a standard normal distribution $N(0, 1)$, and the new forms were assumed to be taken by an equally or more able group with a normal distribution with four levels in mean: 0, 0.02, 0.05, and 0.1, which represent no group difference, two levels of small group differences, and one level of large group difference, respectively. In practice, the group difference is rarely as large as 0.1, this condition was included to represent the worst case scenario.

Evaluation Criteria

The criterion equating relationship was defined as a single group equipercentile equating based on the generating item parameters and two large volume of simulated examinees groups (800,000). The IRT theta abilities of both groups were randomly drawn from the standard normal distribution. Five hundred replications were repeated to yield relatively stable equating results for each condition.

To evaluate the equating results generated from different conditions, overall statistics were calculated to evaluate the amount of error over the entire score scale, including weighted

average absolute bias (WAB), weighted average standard error of equating (WSEE), and weighted average root mean square error (WRMSE), where the weights, $w(x_i)$, were the relative frequency of scores of the new group used in criterion equating to balance out the impact of equating results at extreme scores where little data exists. These three indices were calculated by the following equations:

$$\text{WAB} = \sqrt{\sum_{i=0}^{K} w(x_i)[CAB(x_i)]^2}, \quad (1)$$

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

$$\text{WRMSE} = \sqrt{\sum_{i=0}^{K} w(x_i)[CRMSE(x_i)]^2}, \quad (3)$$

where *K* is the total score of the new form; *i* represents each raw score point; *CAB(x_i)* is conditional absolute bias for score x_i; *CSEE(x_i)* is conditional standard error of equating for score x_i; and *CRMSE(x_i)* is conditional root mean squared error for score x_i. These conditional statistics were calculated using the following formulas:

$$\text{CAB}\ (x_i) = |\bar{e}_Y(x_i)\ -\ e_Y(x_i)|, \quad (4)$$

$$\text{CSEE}\ (x_i) = \sqrt{\frac{\sum_{r=1}^{500}[\hat{e}_{Y,r}(x_i) - \bar{e}_Y(x_i)]^2}{500-1}}, \quad (5)$$

$$\text{CRMSE}(x_i) = \sqrt{CAB(x_i)^2 + CSEE(x_i)^2}, \quad (6)$$

where

$$\bar{e}_Y(x_i)\ = \frac{\sum_{r=1}^{500} \hat{e}_{Y,r}(x_i)}{500}. \quad (7)$$

In Equations 4, 5, and 7, $\hat{e}_{Y,r}(x_i)$ 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.

DTM (Dorans et al., 2003) was used as a criterion to evaluate the magnitude of equating error. Typically, rounded scores were reported in practice. Therefore, if the difference between two unrounded scores is less than half a score unit, the reported rounded scores may be the same. In the current study, DTM of 0.5 was used as a benchmark of acceptable equating error.

Results

Weighted evaluation indices (i.e., WAB, WSEE, and WRMSE) for all investigated conditions are presented in Tables 1 through 3. To better understand the trend of the weighted evaluation indices, graphic presentations of these indices are illustrated in Figures 1 through 10.

Figures 1 through 4 are the results when the assumptions of random groups design were met and there was no group difference. The trends of WAB, WSEE, and WRMSE results for all 16 levels of sample sizes with presmoothing, postsmoothing, and no smoothing, respectively, are shown in Figures 1 through 3. It is obvious that the magnitudes of WSEE and WRMSE decrease as sample size increases. The overall trend of WAB decreases as well, but it is not consistent across different sample sizes. The descending trend is more obvious as sample size increasing from 500 to 3,000, and less obvious as sample size increasing from 3,000 to 8,000. This tendency is consistent across all smoothing methods as shown in Figures 1 through 3. Figure 4 includes all the results shown in Figures 1 to 3 to illustrate the comparison of different smoothing methods on three overall statistics at different levels of sample sizes. As shown in Figure 4, both presmoothing and postsmoothing yield smaller WSEE and WRMSE compared to no smoothing. The lines of WRMSE and WSEE are almost identical, which suggests that WAB doesn't contribute much to WRMSE when there is no group difference. Postsmoothing methods produces slightly smaller WRMSE than presmoothing. Increasing the sample size by 500 yields smaller WRMSE compared to implement of smoothing methods. For example, WRMSE is 0.701 for sample size of 1,500 with no smoothing, which is smaller than WRMSE for sample size of 1,000 with presmoothing (0.799) or postsmoothing (0.784). According to the benchmark of DTM, when sample sizes greater than or equal 3,000 were used for equating, the total error is within acceptable amount.

Figures 5 through 7 present the results when the randomly equivalent groups assumption was violated and plot the overall statistical results for all smoothing methods at different sample sizes the results for group ability difference of 0.02, 0.05, and 0.1, respectively. As we can see from these figures, when group difference increases, the magnitude and trend for WSEE shows very small changes, while the magnitude of WAB increases significantly, which implies that group difference affects systematic error (WAB), but does not affect the random error (WSEE). When group difference is 0.02, WRMSE for sample sizes greater than or equal to 3,500 are

within DTM boundary. However, when group difference equals to 0.05 or 0.1, WRMSE for all studied conditions are outside of DTM boundary.

Figures 8 through 10 show the aggregated effect of each investigation factor. For Figures 8 and 10, the results are aggregated over conditions of group differences of 0, 0.02 and 0.05. The conditions of group difference of 0.1 are not included because they are almost impossible in practice and only for comparison purpose. In general, WRMSE declines as sample size increases. The declining rate is relatively rapid for sample sizes from 500 to 3,000, and is relatively slow for sample sizes from 3,000 to 8,000 (see Figure 8). The DTM criterion suggests that equating results with sample sizes greater than or equal 3,000 produce almost acceptable level of equating accuracy (see Figure 8). WAB and WRMSE increase as group difference increases, but WSEE doesn't change (see Figure 9). Smoothing produces smaller WRMSE and thus improves equating accuracy (see Figure 10).

Discussion

The primary goal of this study was to investigate the minimum sample size needed for equipercentile equating to produce adequate equating results for the RG design. The impacts of smoothing methods and the possibility of existence of group ability difference when the assumption of randomly equivalent groups is violated are also considered. Data were generated using simulation techniques based on AP US History examinations.

Overall, larger sample sizes produced more accurate equating results. The extent of growth in equating accuracy is more significant for sample sizes from 500 to 3,000, and less significant for sample sizes from 3,000 to 8,000. In agreement with previous research, smoothing yielded more accurate equating. When proper procedures were followed and the requirements of randomly equivalent groups assumption for the random group design were met, a sample size of 3,000 is adequate to yield equating results with an acceptable accuracy. When the randomly equivalent groups assumption is violated to certain extent (group ability difference larger than 0.05), the equating results would be bias no matter how large the sample size were used. When the sample size can be used is limited, implementing a smoothing method would improve equating accuracy.

A few limitations of this study should be recognized. First, because this is a simulation study, it should be cautious when interpreting the results. Second, all simulated data are based on the AP US History Examinations, it is worthwhile to verify the findings with other datasets.

Future research could further investigate different equating methods, generating IRT models, or different equating situations with mixed-format tests of different lengths.

References

Asiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory and Practice, 16(2),* 647-668.

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.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph Number 1). Iowa City, IA: CASMA, The University of Iowa.

Cai, L. (2015). *flexMIRT* (Version 3.0) [Computer Program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46*(2), 135–158.

Dorans, N. J. (Ed.). (2003). *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS Research Report RR-03-27). Princeton, NJ: Educational Testing Service.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* (Version 1.0) [Computer Program]. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

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

Liu, C., & Kolen, M. J. (2011a). Evaluating smoothing in equipercentile equating using fixed smoothing parameters. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 1)*. (CASMA Monograph Number 2.1) (pp. 213-236). Iowa City, IA: CASMA, The University of Iowa.

Liu, C., & Kolen, M. J. (2011b). Automated selection of smoothing parameters in equipercentile equating. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 1)*. (CASMA Monograph Number 2.1) (pp. 237–261). Iowa City, IA: CASMA, The University of Iowa.

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

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42(4),* 309-330.

Table 1

Weighted Average Absolute Bias for Each Condition

Sample Size	Group Difference and Smoothing Method											
	0			0.02			0.05			0.1		
	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO
500	0.281	0.274	0.303	0.205	0.195	0.235	0.467	0.462	0.484	1.048	1.044	1.059
1000	0.154	0.154	0.168	0.357	0.355	0.363	0.776	0.775	0.778	1.325	1.323	1.329
1500	0.097	0.087	0.103	0.289	0.286	0.293	0.640	0.638	0.644	1.268	1.266	1.270
2000	0.048	0.044	0.064	0.209	0.206	0.211	0.548	0.546	0.550	1.171	1.170	1.173
2500	0.103	0.100	0.107	0.141	0.137	0.145	0.449	0.447	0.450	1.055	1.053	1.056
3000	0.100	0.102	0.105	0.167	0.165	0.169	0.501	0.499	0.502	1.085	1.083	1.086
3500	0.073	0.075	0.082	0.110	0.109	0.115	0.471	0.470	0.472	1.072	1.071	1.073
4000	0.070	0.071	0.082	0.154	0.153	0.157	0.526	0.524	0.527	1.112	1.110	1.112
4500	0.047	0.051	0.056	0.224	0.224	0.225	0.568	0.567	0.569	1.162	1.161	1.163
5000	0.084	0.083	0.088	0.171	0.173	0.174	0.531	0.531	0.532	1.113	1.113	1.114
5500	0.063	0.064	0.069	0.240	0.241	0.243	0.587	0.587	0.588	1.171	1.170	1.171
6000	0.056	0.057	0.062	0.238	0.237	0.239	0.592	0.592	0.593	1.189	1.188	1.190
6500	0.051	0.052	0.055	0.245	0.245	0.246	0.619	0.618	0.619	1.200	1.199	1.200
7000	0.059	0.058	0.064	0.228	0.227	0.229	0.594	0.594	0.595	1.186	1.185	1.186
7500	0.032	0.033	0.036	0.217	0.216	0.219	0.591	0.590	0.592	1.184	1.183	1.184
8000	0.079	0.082	0.083	0.223	0.223	0.225	0.584	0.583	0.585	1.168	1.168	1.169
Average	0.087	0.087	0.096	0.214	0.212	0.218	0.565	0.564	0.568	1.157	1.155	1.158

Table 2

Weighted Standard Error of Equating for Each Condition

Sample Size	Group Difference and Smoothing Method											
	0			0.02			0.05			0.1		
	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO
500	1.062	1.050	1.146	1.067	1.054	1.152	1.086	1.076	1.176	1.081	1.069	1.171
1000	0.800	0.785	0.855	0.770	0.755	0.826	0.776	0.762	0.833	0.803	0.785	0.855
1500	0.639	0.625	0.683	0.645	0.630	0.688	0.650	0.633	0.691	0.646	0.628	0.689
2000	0.550	0.533	0.586	0.560	0.545	0.595	0.554	0.538	0.590	0.568	0.553	0.604
2500	0.495	0.481	0.527	0.501	0.487	0.532	0.498	0.484	0.530	0.504	0.490	0.536
3000	0.445	0.432	0.476	0.461	0.448	0.490	0.465	0.451	0.492	0.473	0.460	0.501
3500	0.422	0.411	0.447	0.442	0.429	0.466	0.426	0.412	0.451	0.436	0.425	0.464
4000	0.403	0.391	0.425	0.398	0.386	0.422	0.405	0.395	0.430	0.420	0.409	0.444
4500	0.383	0.372	0.405	0.382	0.371	0.404	0.391	0.380	0.413	0.410	0.399	0.430
5000	0.371	0.362	0.392	0.366	0.357	0.387	0.380	0.371	0.401	0.386	0.376	0.406
5500	0.350	0.341	0.370	0.361	0.352	0.380	0.360	0.351	0.379	0.383	0.374	0.400
6000	0.336	0.329	0.355	0.343	0.334	0.360	0.349	0.340	0.367	0.367	0.359	0.385
6500	0.330	0.321	0.347	0.340	0.332	0.357	0.338	0.330	0.355	0.360	0.352	0.377
7000	0.320	0.312	0.337	0.319	0.312	0.337	0.338	0.330	0.354	0.345	0.337	0.361
7500	0.314	0.306	0.329	0.314	0.307	0.330	0.322	0.315	0.338	0.352	0.345	0.366
8000	0.304	0.296	0.319	0.308	0.300	0.322	0.316	0.309	0.330	0.328	0.321	0.343
Average	0.470	0.459	0.500	0.474	0.462	0.503	0.478	0.467	0.508	0.491	0.480	0.521

Table 3

Weighted Average Root Mean Square Error for Each Condition

Sample Size	Group Difference and Smoothing Method											
	0			0.02			0.05			0.1		
	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO	PRE	POST	NO
500	1.099	1.085	1.185	1.087	1.072	1.176	1.182	1.171	1.272	1.506	1.494	1.579
1000	0.815	0.800	0.872	0.849	0.834	0.902	1.097	1.087	1.140	1.549	1.539	1.580
1500	0.646	0.631	0.690	0.707	0.692	0.748	0.912	0.899	0.944	1.424	1.414	1.445
2000	0.552	0.535	0.589	0.598	0.583	0.631	0.779	0.767	0.807	1.302	1.294	1.319
2500	0.505	0.492	0.537	0.520	0.505	0.552	0.671	0.659	0.695	1.169	1.161	1.185
3000	0.456	0.444	0.487	0.490	0.478	0.518	0.683	0.673	0.703	1.184	1.177	1.196
3500	0.428	0.417	0.455	0.455	0.442	0.480	0.635	0.625	0.653	1.157	1.152	1.169
4000	0.409	0.398	0.433	0.426	0.415	0.450	0.664	0.656	0.680	1.188	1.183	1.197
4500	0.386	0.376	0.409	0.442	0.433	0.462	0.690	0.683	0.703	1.232	1.227	1.240
5000	0.381	0.372	0.402	0.404	0.397	0.425	0.653	0.648	0.666	1.178	1.174	1.186
5500	0.356	0.347	0.376	0.434	0.426	0.451	0.689	0.683	0.700	1.232	1.228	1.238
6000	0.341	0.334	0.360	0.417	0.409	0.432	0.688	0.683	0.697	1.245	1.241	1.250
6500	0.334	0.325	0.351	0.419	0.412	0.434	0.705	0.701	0.714	1.253	1.250	1.258
7000	0.326	0.318	0.343	0.392	0.386	0.407	0.684	0.679	0.692	1.235	1.232	1.240
7500	0.315	0.308	0.331	0.382	0.375	0.396	0.673	0.669	0.681	1.235	1.232	1.239
8000	0.314	0.308	0.329	0.380	0.374	0.393	0.664	0.660	0.672	1.213	1.211	1.218
Average	0.479	0.468	0.509	0.525	0.515	0.554	0.754	0.746	0.776	1.269	1.263	1.284

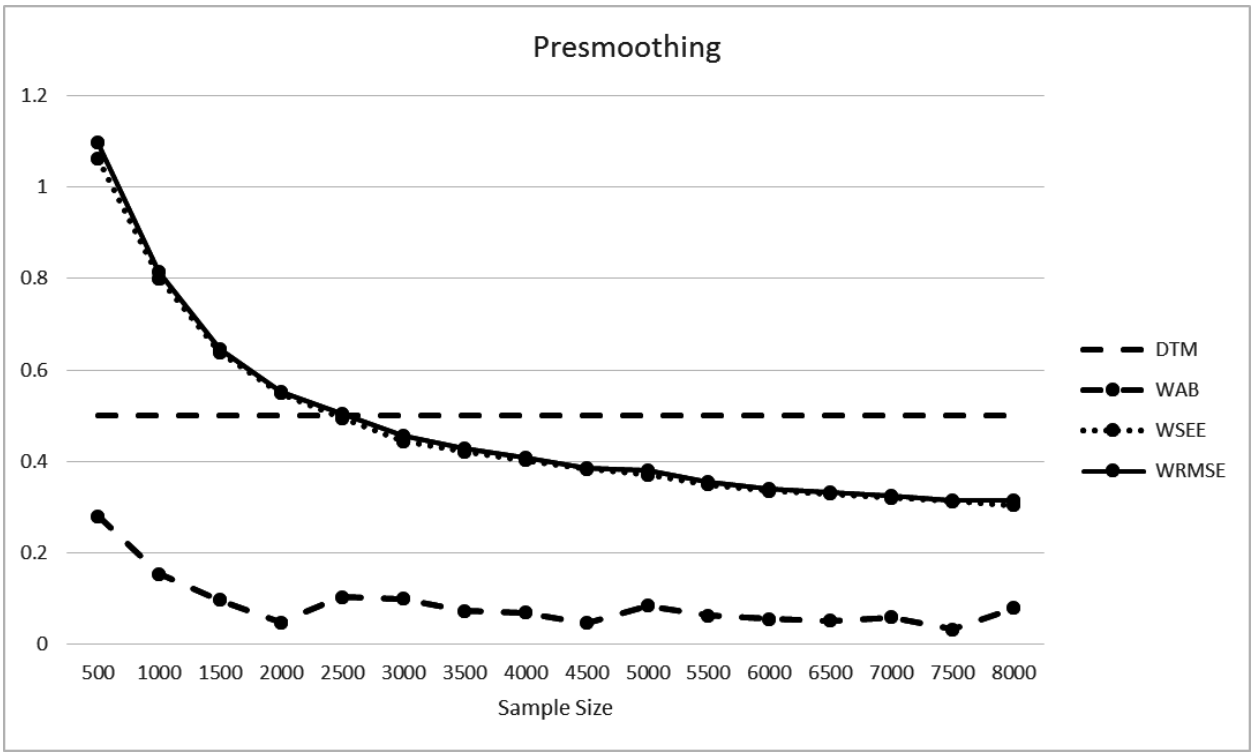


Figure 1. Weighted evaluation indices for different sample sizes when presmoothing was used and group difference was 0.

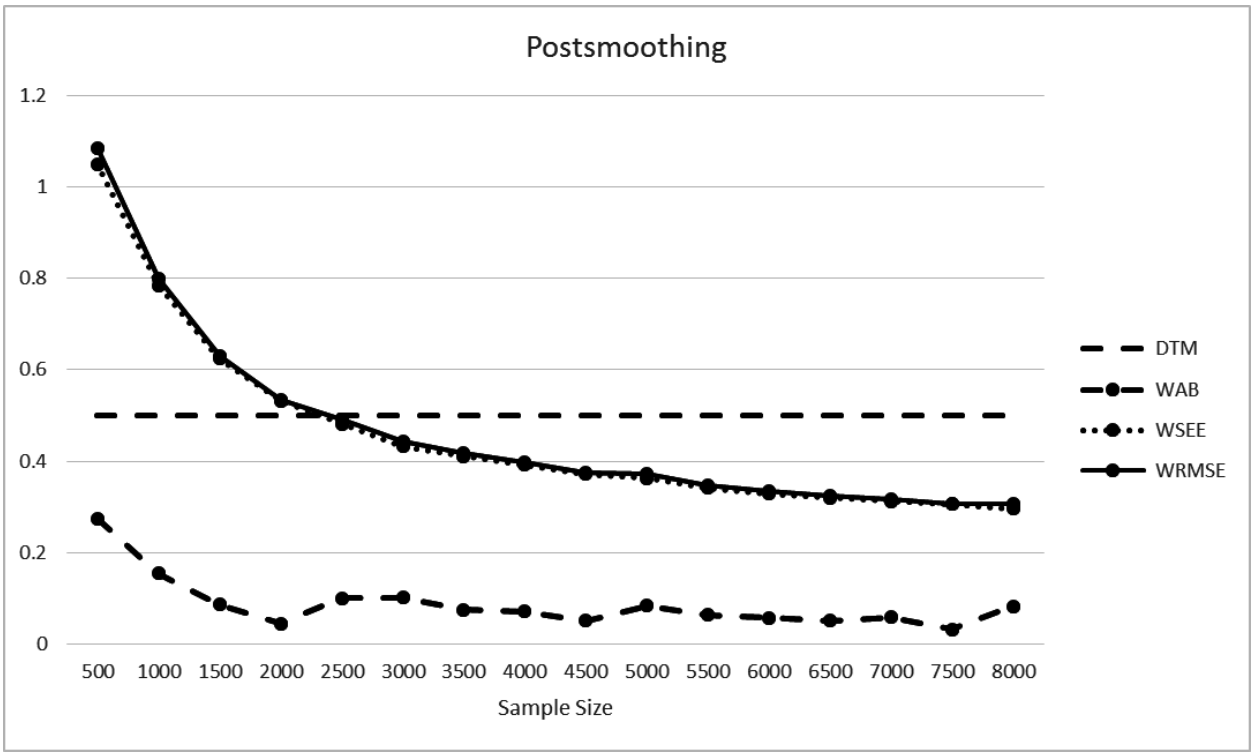


Figure 2. Weighted evaluation indices for different sample sizes when postsmoothing was used and group difference was 0.

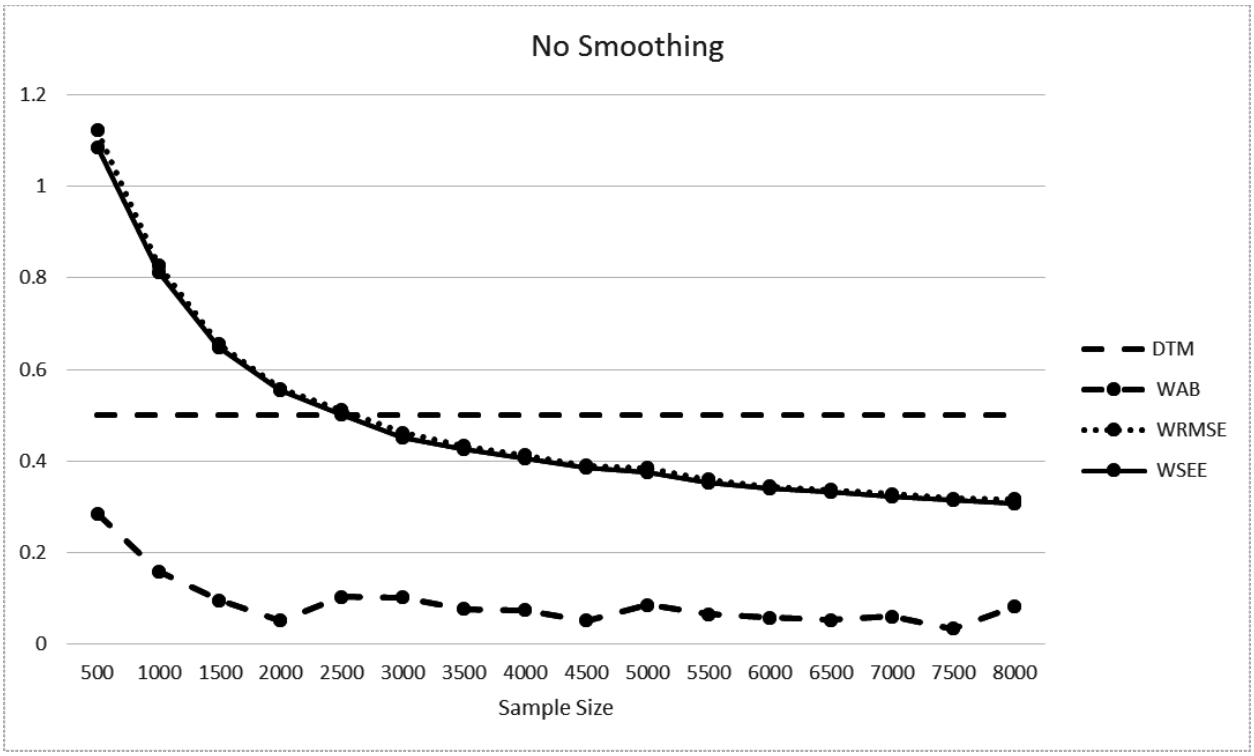


Figure 3. Weighted evaluation indices for different sample sizes when no smoothing was used and group difference was 0.

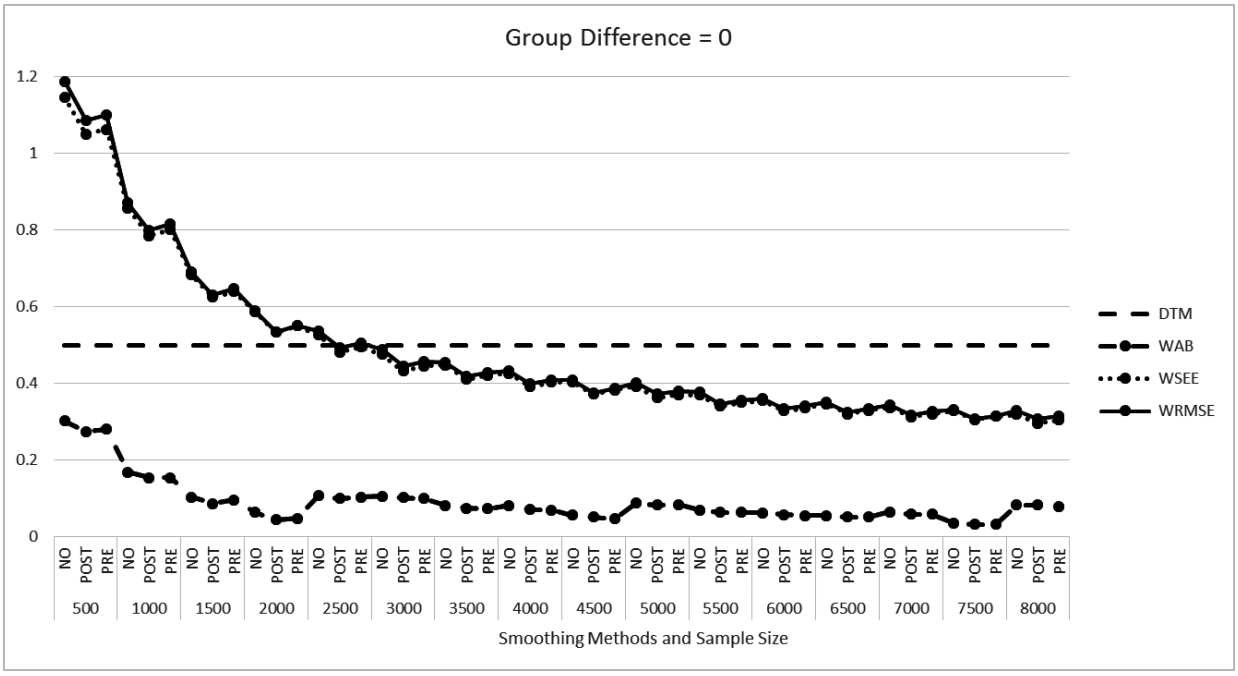


Figure 4. Weighted evaluation indices for different sample sizes when all smoothing methods were presented and group difference was 0.

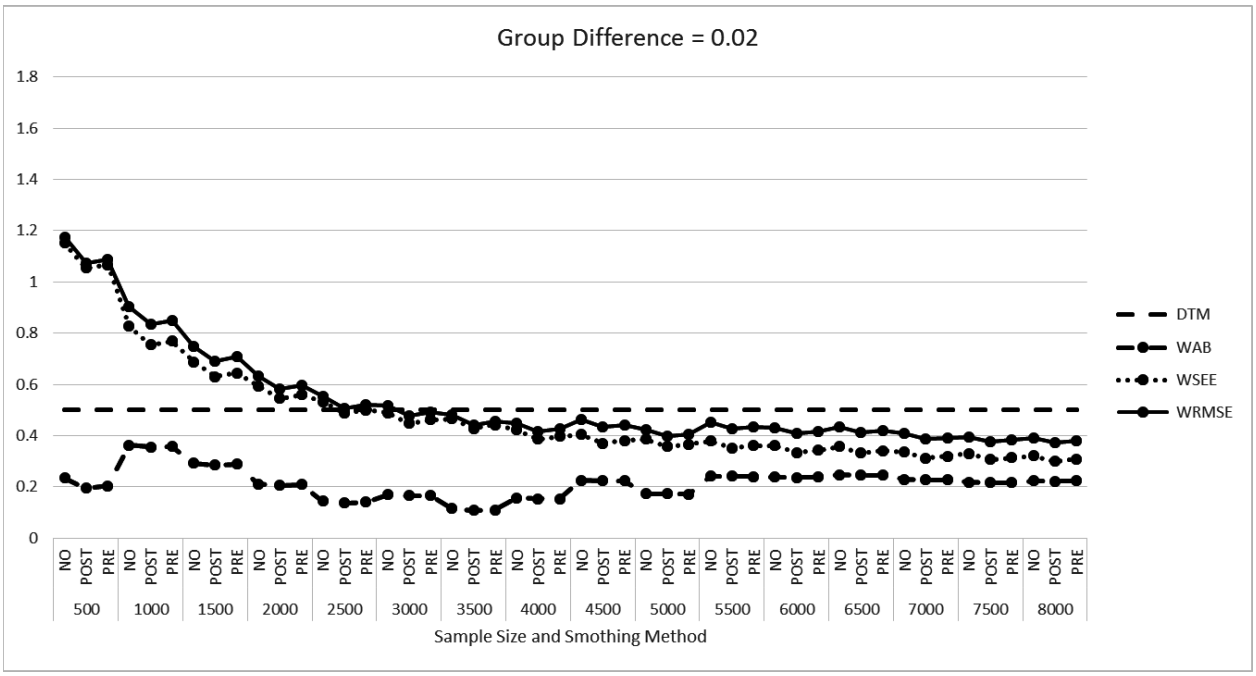


Figure 5. Weighted evaluation indices for different sample sizes when all smoothing methods were presented and group difference was 0.02.

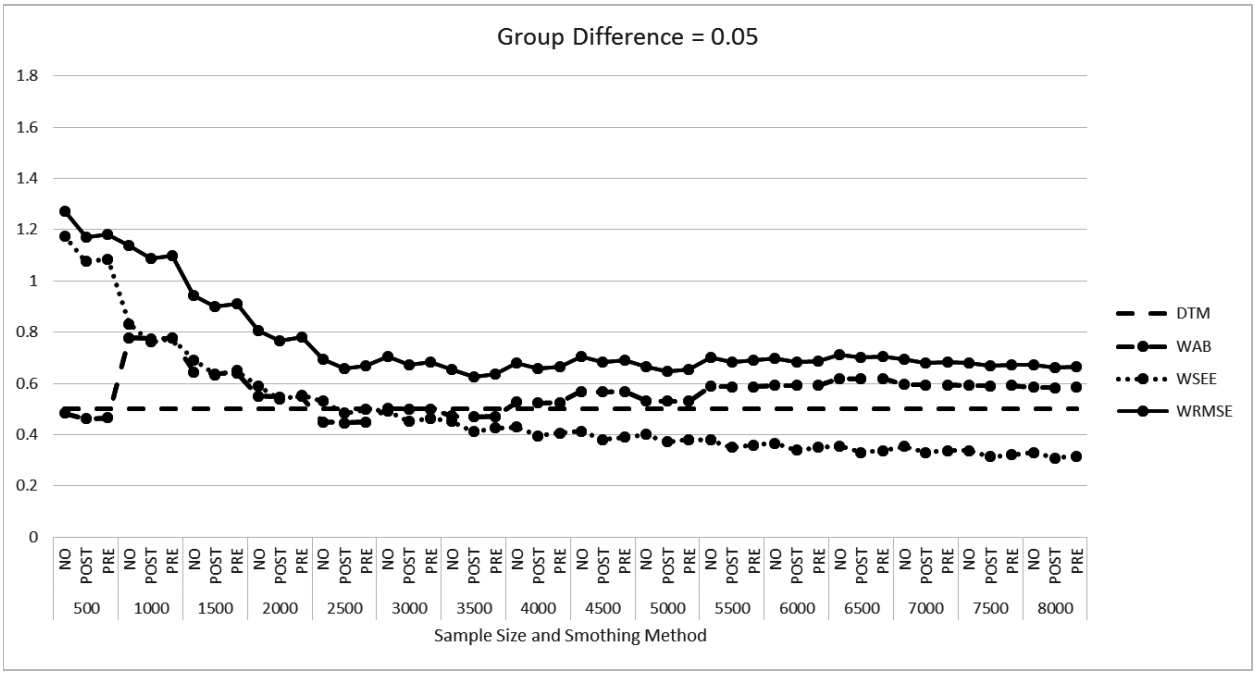


Figure 6. Weighted evaluation indices for different sample sizes when all smoothing methods were presented and group difference was 0.05.

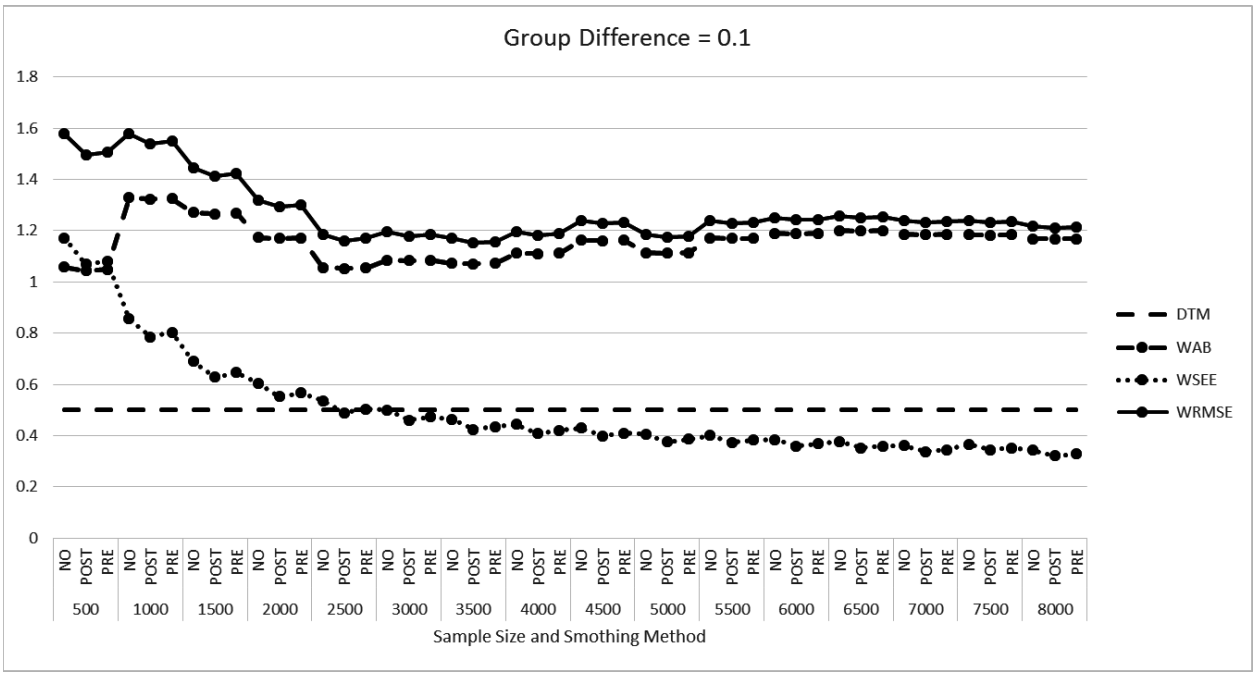


Figure 7. Weighted evaluation indices for different sample sizes when all smoothing methods were presented and group difference was 0.1.

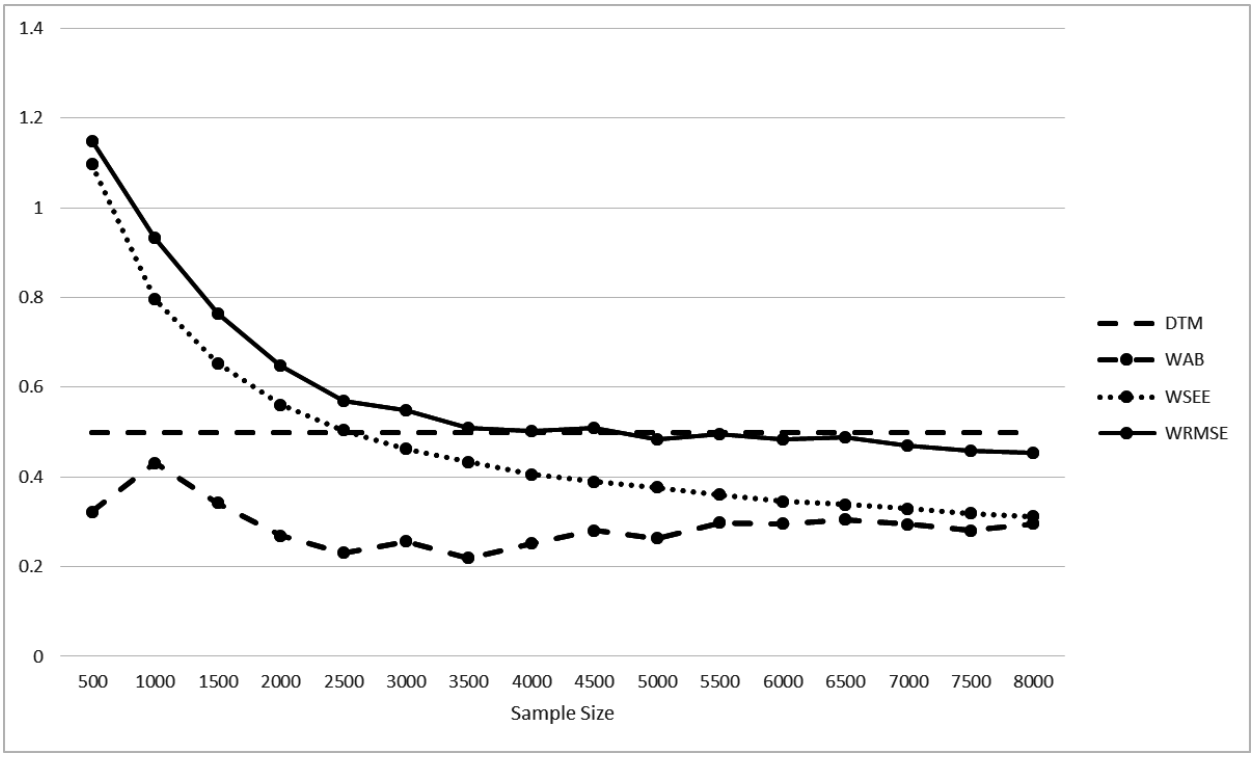


Figure 8. Aggregated sample size effect for conditions with group differences of 0, 0.02 and 0.05.

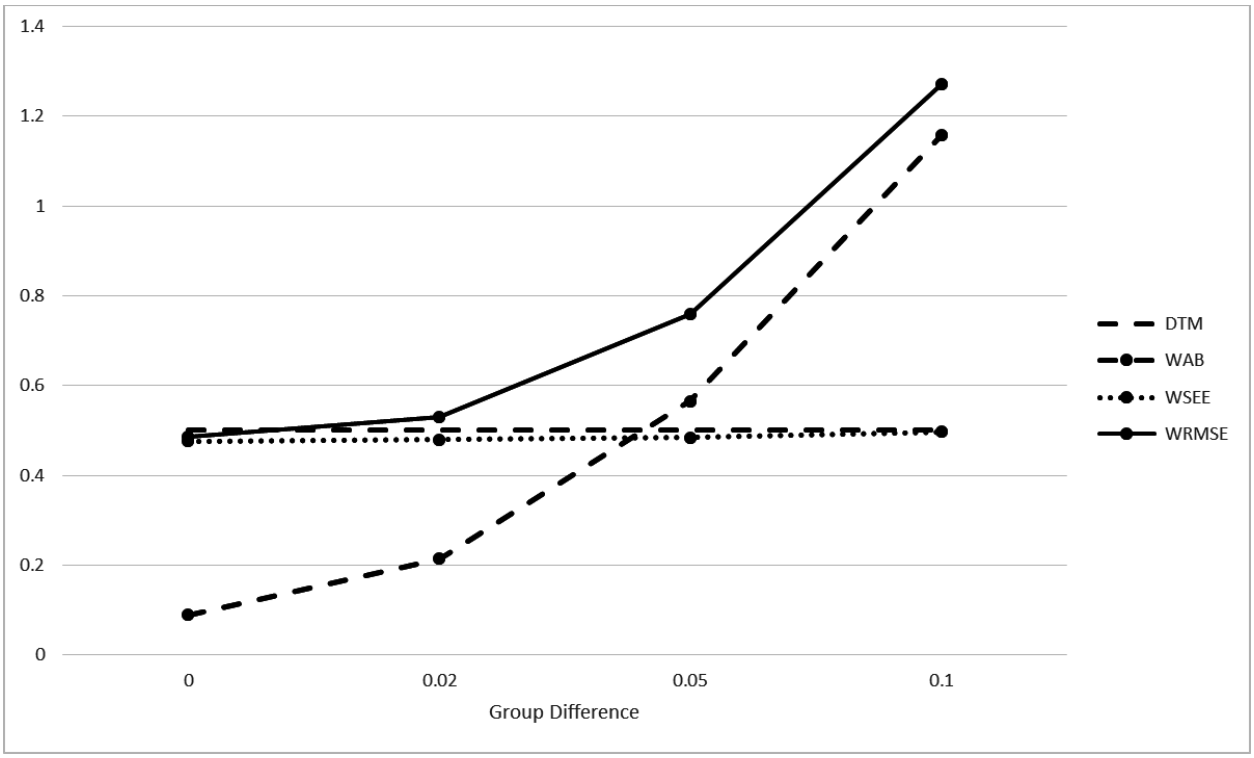


Figure 9. Aggregated group difference effect for all the conditions.

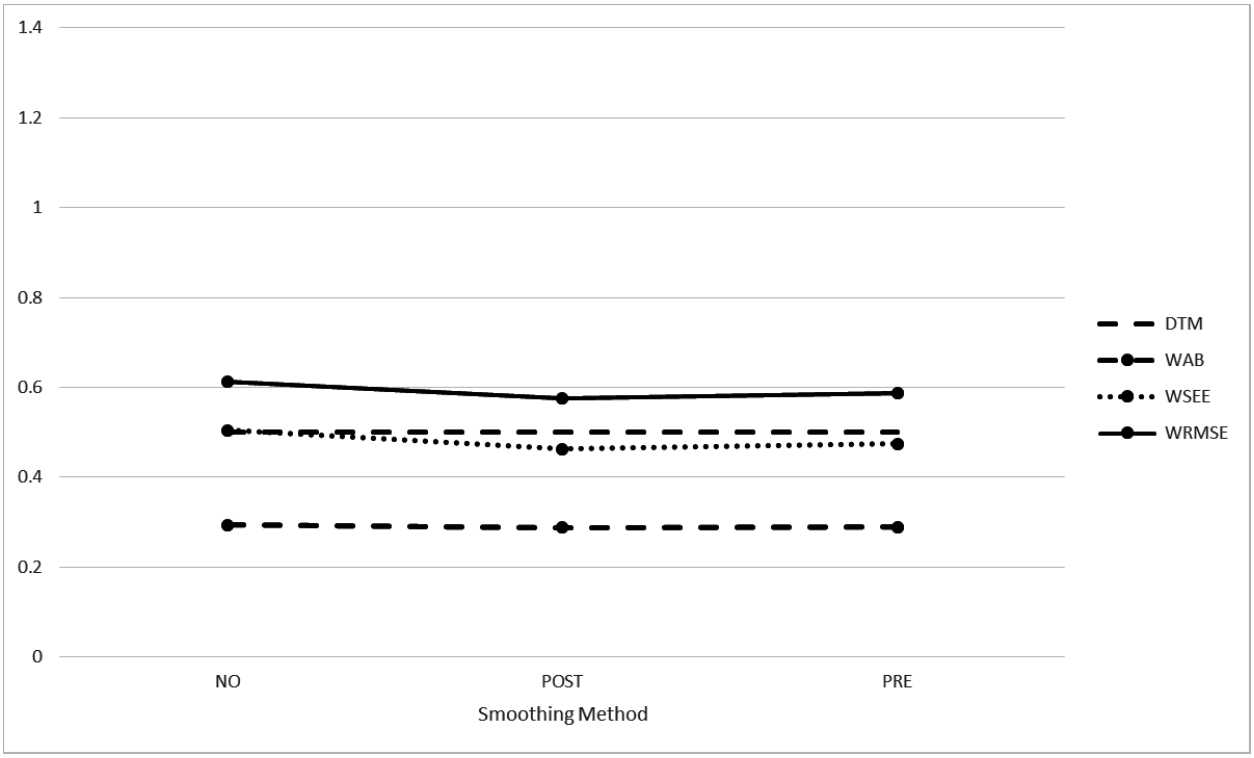


Figure 10. Aggregated smoothing method effect for conditions with group differences of 0, 0.02 and 0.05.

Chapter 5: Simple-Structure MIRT True-Score Equating for Mixed-Format Tests

Stella Y. Kim and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

This study proposes a true-score equating procedure for mixed-format tests under the simple-structure multidimensional item response theory (SS-MIRT) framework. Based on real data analysis, the proposed equating procedure was empirically compared with four competing procedures: equipercentile equating with presmoothing, unidimensional IRT (UIRT) observed-score equating, UIRT true-score equating, and SS-MIRT observed-score equating procedures. In addition, this study evaluated the performance of the SS-MIRT true-score equating procedure through a simulation study. The results from real data analysis reveal that the proposed SS-MIRT true-score equating procedure behaves similarly to other IRT-based procedures including UIRT observed-score and true-score equating. The simulation study results show that the proposed equating procedure consistently outperforms the UIRT true-score equating procedure. More specifically, when multidimensionality occurs substantially, a larger bias is observed for the UIRT true-score equating procedure whereas the proposed procedure is relatively robust to the occurrence of multidimensionality of a test.

Simple-Structure MIRT True-Score Equating for Mixed-Format Tests

With the growing popularity of mixed-format tests, it has become increasingly important to provide an adequate basis for analyzing the psychometric properties of such tests. The traditional unidimensional item response theory (UIRT) framework often fails to meet the emerging need for a psychometric model that can precisely model the structure of mixed-format tests, because of its strong assumptions about the data structure. Particularly, the essence of the UIRT model rests on the unidimensionality assumption, which dictates that a test only measures a single ability. However, this assumption is not likely to be fulfilled for mixed-format tests because the two types of item format – multiple-choice (MC) and free-response (FR) items – might measure distinct, but still correlated abilities. In this case, the simple-structure multidimensional item response theory (SS-MIRT) framework might adequately reflect the underlying structure of such mixed-format data.

Recognition of the strengths of the SS-MIRT model in taking multiple abilities into account has led to the use of this model in numerous research studies. For example, the SS-MIRT approach has been acknowledged as one potential way to analyze a test composed of multiple test batteries, under the assumption that items in each subtest are associated with a single subtest-specific ability (Cheng, Wang, & Ho; 2009; de la Torre, 2008; de la Torre & Patz, 2005; Wang, Chen, & Cheng, 2004). The previous findings show that the SS-MIRT model substantially improves measurement precision, especially when each subtest is short and the test battery contains a large number of tests. The SS-MIRT model has also been discussed in the context of a computerized adaptive test (Li & Schafer, 2005; Segall, 1996; Wang & Chen, 2004). Previous studies have reported that adopting a multidimensional computerized adaptive test algorithm using the SS-MIRT model improves efficiency in measuring multiple content coverage areas and provides more accurate ability estimates than a UIRT model.

In the context of equating, when the assumptions of the UIRT models are not met, the inaccuracy of the resultant equating relationships under the UIRT framework is threatened. Obviously, the use of an equating procedure that leads to unacceptably inaccurate results should be avoided. As an alternative approach, there has been a growing demand for equating procedures that are based on the multidimensional item response theory (MIRT) framework (Brossman & Lee, 2013; Lee & Brossman, 2012; Lee & Lee, 2016). The equating procedures that have been developed under the MIRT framework include observed-score equating and true-score equating with a multidimensional two parameter logistic model (Brossman & Lee,

2013), an observed-score equating procedure using a bi-factor MIRT model (Lee & Lee, 2016), a full MIRT observed-score equating procedure (Peterson & Lee, 2014), and an observed-score equating procedure with the SS-MIRT model (Lee & Brossman, 2012). Lee and Brossman (2012) found that the SS-MIRT observed-score equating procedure produced more accurate equating results than the UIRT procedure when the data were multidimensional.

Compared to other MIRT approaches, SS-MIRT equating has several compelling features (Lee & Brossman, 2012). One promising characteristic of this approach is its calibration efficiency. The calibration process under the SS-MIRT framework is relatively simple and faster than other MIRT models because each item loads only on one ability, whereas most MIRT models relate multiple abilities to a single item. Also, SS-MIRT allows for a straightforward interpretation of the data structure such as the relationships among dimensions and the weights for each dimension.

In spite of such useful features, only a limited number of studies have dealt with the SS-MIRT approach in the context of equating (Lee & Brossman, 2012). In particular, a true-score equating procedure under the SS-MIRT framework has not been developed in the literature. Therefore, the primary goal of this study is to propose a true-score equating procedure under the SS-MIRT framework for mixed-format tests. The application of the proposed equating procedure is described using real data analysis. In addition, a simulation study is conducted to further investigate the performance of the SS-MIRT true-score (SMT) equating procedure.

SS-MIRT True-Score Equating Procedure

In traditional UIRT true-score equating, it is assumed that the true score on one form is equivalent to the true score on another form for a given ability θ, as long as the item parameters are on the same scale. In such a univariate case, the test characteristic curves (TCC) for each of the two forms are used to relate IRT ability to true score.

The SS-MIRT model allows for multiple abilities and, consequently, an examinee's expected number-correct score is conditional on the θ-vector. A test characteristic surface (TCS) represents the relationship between the expected number-correct scores and θ-vector. The challenge of conducting true-score equating using multiple ability dimensions comes from the fact that there is no unique combination of thetas that corresponds to a particular true score. To deal with such complexities, a procedure is proposed in this paper that does not involve direct use of the multidimensional space. Instead, the multidimensional space is

collapsed into a set of unidimensional spaces, in which each composite score is a linear combination of all possible univariate components corresponding to the same true score.

The proposed SMT equating procedure involves the following steps.

(1) Calibrate MC and FR items on each form separately using the SS-MIRT models.

(2) Conduct standard UIRT true-score equating for the MC and FR sections separately.

(3) Compute the equated composite score for each pair of MC and FR scores through a weighted sum of the MC and FR equated scores. Note that the composite score, x, is a weighted sum of two section scores rounded to the nearest integer. That is, $X = [w_M X_M + w_F X_F]$, where w_M and w_F represent pre-specified weights for MC and FR sections, respectively; and the brackets, [ ], are used to denote the rounding operation. Then, the equated composite score for each pair of x_M and x_F can be expressed as

$$eq(X = x|x_M, x_F) = w_M eq_M(x_M) + \ w_F eq_F(x_F), \tag{1}$$

where $eq_M(x_M)$ and $eq_F(x_F)$ denote the typical UIRT true-score equating equivalent for score x_M for MC section, and for score x_F for FR section, respectively. The section scores, x_M and x_F, are typically number-correct integer scores, but not necessarily.

(4) Estimate a bivariate frequency distribution for the MC and FR scores conditional on each composite score $f(x_M, x_F|x)$ for new form.

(5) Compute the final equated score, which is a weighted sum of equated composite scores. The summation is taken over all possible pairs of MC and FR scores that lead to a particular composite score *x*, and the weight, $f(x_M, x_F|x)$, is the relative frequency of the specific MC and FR score pair conditioning on a certain composite score. Final equated score $eq(x)$ is defined by the following equation:

$$eq(x) = \ \sum_{X=[w_M X_M + w_F X_F]} f(x_M, x_F|x)\big(w_M eq_M(x_M) + \ w_F eq_F(x_F)\big). \tag{2}$$

Note that this equating procedure can be used for any equating design as long as all of the parameter estimates are on the same scale.

The proposed SMT equating procedure is based on the following assumptions: (a) each item measures only a single ability; (b) the underlying multiple abilities are allowed to be correlated with each other; (c) a cluster of items measuring the same ability can be modeled accurately using a UIRT model; (d)

ct; and (e) correlations between abilities are captured and reflected in the equating process by applying frequency weights to compute the final equated score, where the frequency weights are a multivariate (bivariate with MC and FR sections)

observed-score distribution. Since there are multiple combinations of MC and FR scores that will produce the same composite score, our goal here is to define a single equated composite score that is most representative of all the combinations. In order to achieve this goal, the frequency of each combination is used, which also reflects a correlation between two section scores. It is important to note that correlations between abilities are not necessarily the same for the two groups as long as the two forms measure the same construct for each section.

Estimating Joint Bivariate Score Distributions

Conducting SMT equating requires the use of a discrete bivariate score frequency distribution for MC and FR scores. In this paper, three methods for estimating the bivariate score distributions are suggested, and compared in terms of their precision which are provided in the Result section. Once the bivariate distribution is obtained using one of the three proposed methods described below, a set of weights $f(x_M, x_F|x)$ in Equation 1 can be computed for each composite score point.

Actual distribution. The simplest approach for obtaining the bivariate frequency distribution is to use the actual score distribution by creating a contingency table for MC and FR scores. A bivariate distribution between the observed MC scores and the observed FR scores can be easily obtained from the data at hand.

One practical concern involved in this method is that there is the potential for non-consecutiveness in a conversion table. Due to the random errors, the observed distribution is usually bumpy, potentially leading to zero frequency at some score points. If all weights are given to a specific score pair in Equation 1, rather than considering all possible score pairs, the subsequent equating results may be inaccurate, allowing the random error components to largely determine the equating relationships. This may also result in a non-consecutive conversion table, meaning that an equated score that corresponds to a smaller raw score point is larger than an equated score that corresponds to a larger score point.

Log-linear smoothed distribution. Instead of using an observed-score distribution itself, a bivariate log-linear model can be used to attain a smoothed joint frequency distribution. Smoothing usually reduces random errors involved in an observed-score distribution. Equation 3 provides a log-linear model for a test having a possible MC score range of $z_1, z_2, \dots, z_k$ and a possible FR score range of $y_1, y_2, \dots, y_l$ (Moses & Holland, 2010). The log-linear model has the form of the log of the expected score probabilities (p_{kl}) in terms of a polynomial function of two section scores, which can be expressed as

$$log_e(p_{kl}) = \beta_0 + \sum_{a=1}^{A} \beta_a \, z_k^a + \sum_{b=1}^{B} \beta_{A+b} \, y_l^b + \sum_{c=1}^{C} \sum_{d=1}^{D} \beta_{cd} \, z_k^c y_l^d, \quad (3)$$

where β_0 is a normalizing constant, the z_k^a and y_l^b are a function of MC and FR section scores, and β is the parameter to be estimated. The number of moments for the univariate fitted distribution is determined by the values of *A* and *B*. Likewise, the number of moments of the bivariate smoothed distribution is determined by the values of *C* and *D*.

IRT-fitted distribution. Another approach to estimating the bivariate score distribution is to produce a model-fitted bivariate distribution using the SS-MIRT model. Based on the assumption of conditional independence with respect to two ability dimensions, the probability of a correct response to an item for a particular examinee are mutually independent of the probabilities for other items, conditioning on examinee's abilities θ_M and θ_F. Thus, given item parameters, a conditional bivariate score distribution for each pair of section abilities can be determined as the product of each conditional observed score distribution. This can be expressed as:

$$f(X_M = x_M, X_F = x_F|\theta_M,\theta_F) = f(X_M = x_M|\theta_M) f(X_F = x_F|\theta_F), \quad (4)$$

where $f(x_M|\theta_M)$ and $f(x_F|\theta_F)$ represent the conditional distribution of getting a score of x_M and x_F for an examinee with abilities θ_M and θ_F, respectively. The conditional distribution for each section can be produced using Lord and Wingersky (1984) formula for MC items or using a modified version of the Lord-Wingersky formula by Hanson (1994) for FR items.

Finally, the bivariate score distribution can be obtained by aggregating conditional bivariate score distributions over all pairs of two latent abilities, $g(\theta_M, \theta_F)$, denoted as:

$$f(x_M, x_F) = \int_{\theta_M} \int_{\theta_F} f(x_M|\theta_M) f(x_F|\theta_F) g(\theta_M, \theta_F) d\theta_M d\theta_F. \quad (5)$$

The joint bivariate distributions can be approximated by replacing integrals in Equation 5 with summations, which defines the marginal distribution by the following formula:

$$f(x_M, x_F) = \sum_{\theta_M} \sum_{\theta_F} f(x_M|\theta_M) f(x_F|\theta_F) Q(\theta_M, \theta_F), \quad (6)$$

where $Q(\theta_M, \theta_F)$ is the discrete quadrature density function of the bivariate ability distribution.

Once the joint bivariate distribution is found, the last step is to obtain the bivariate frequency distribution for the MC and FR scores conditional on each composite score, which can be shown as

$$f(x_M, x_F|x) = \frac{f(x_M, x_F)}{\sum_{X=[w_M X_M + w_F X_F]} f(x_M, x_F)}. \quad (7)$$

An Illustrative Example

Real data analyses were conducted to evaluate the feasibility and applicability of the SMT equating procedure compared to other competing procedures.

Data

Two forms of the Advanced Placement (AP) English Language exam, administered in 2011 and 2013, were used. For illustrative purposes, some aspects of the data were arbitrarily manipulated: (a) the section weights for MC and FR items were set to 1:1 to reduce the computational complexity, (b) the random groups design was used although the data ware originally collected under the common-item nonequivalent groups design (CINEG), (c) 6,000 examinees were sampled from the original data for each form, and (d) normalized scale scores ranging from 0 to 70 were constructed for the raw-to-scale score conversion. Such modifications were employed because the primary aim of this study was to explore the performance of the equating procedures, not to examine the psychometric properties of the AP exams used.

After weights were applied (i.e., a simple sum of the MC and FR section scores), the maximum composite raw scores were equal to 81 and 82 for the old and new forms, respectively. The old form contained 54 MC items scored 0-1 and 3 FR items scored 0-9, and the new form consisted of 55 MC items scored 0-1 and 3 FR items scored 0-9. The estimated classical disattenuated correlation between section scores was .82 and .80 for the new and old forms, respectively, using coefficient alpha as a reliability estimate for each section.

As mentioned earlier, this study was conducted under the random groups design. Although in an operational setting the old and new forms share a set of common items, they were treated as unique items for each form. Also, since the data were considered as if they were collected under the random groups design, it was necessary to verify that the ability levels were approximately equivalent between the two sampled groups. As a means of such data verification, the standardized mean difference (Dorans, 2000) was computed based on common item scores. For the data used in this study, the effect size turned out to be -.019. This indicates that examinees taking the old form were slightly higher achieving than those

taking the new form, but not to a great extent. Therefore, the CI effect size of -.019 was considered small enough to be able to use the random groups design. Note again that because the purpose of the real data analysis is to demonstrate how the proposed SMT equating method behaves relative to other methods, the confounding effect of data collection design may not be of a primary concern.

Analysis

For SMT, two sets of equating relationships were obtained for the MC and FR sections separately by conducting UIRT true-score equating using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). The three-parameter logistic (3PL) model (Birnbaum, 1968) and the graded response (GR) model (Samejima, 1969) were used for the MC and FR items, respectively. Finally, a bivariate joint score distribution was found using the IRT-fitted distribution method. The IRT-fitted distribution method was chosen among the three methods because it is most consistent with the IRT equating procedure itself. A detail comparison among the proposed methods was made through the simulation study presented in the following section.

A model-based bivariate score distribution was found using a computer program R (R Core Team, 2014). First, a bivariate normal ability distribution was defined, which is denoted as $BN(0,0,1,1,\hat{\rho}_{\theta_M\theta_F})$. Note that the correlation between the two section abilities was estimated for each replication, at the time when item parameters were calibrated under the SS-MIRT framework using *flexMIRT* (Cai, 2017). Then, the estimated correlation for the new form was regarded as a correlation for the population ability distribution. Second, the conditional bivariate observed score distribution at each combination of θ_M (MC ability) and θ_F (FR ability) was obtained using the recursive formulas by Lord and Wingersky (1984) for MC items and Hanson (1994) for FR items. Last, the bivariate observed score distribution was found by aggregating conditional bivariate score distributions across the bivariate theta distribution. For the bivariate theta distribution, 41 × 41 bivariate quadrature points and weights were used with a theta range of -4 ~ +4 for each ability.

The results of the proposed procedure were compared to the results for (a) traditional equipercentile with log-linear presmoothing with a degree of 6 (EQ), (b) UIRT true-score (UT), (c) UIRT observed-score (UO), and (d) SS-MIRT observed-score (SMO) equating procedures. The equating relationships for the first three procedures were found using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009) and for SMO, the equating relationship was obtained using a program written for this research purpose.

It should be noted that prior to IRT equating, it is necessary to put item and ability parameters for both forms on the same scale, which is known as a scale linking process. However, the random groups design employed in this study eliminated the necessity for scale linking procedure, given that two groups are assumed to be essentially equivalent in ability. It is important to emphasize that the use of the random groups design in this paper does not suggest that SMT equating is only applicable for this specific data collection design. SMT equating can be performed under the CINEG design once the parameters are transformed to be on the same scale. In this case, various MIRT linking procedures (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2009; Thompson, Nering, & Davey, 1997; Yon, 2006) can be used to achieve the scale transformation.

Results of the Illustrative Example

Before conducting equating, descriptive statistics for the new and old form groups were inspected and are summarized in Table 1. As mentioned earlier, the classical disattenuated correlation for the new form was slightly higher than that of the old form, although both forms showed moderate levels of correlation. Note that in real data analysis, there is no direct way to quantify the precision of each equating procedure, because the true equating relationship is unknown. Thus, the primary focus of this section was to compare the SMT procedure with various other equating procedures and to examine if the proposed procedure behaves reasonably relative to the competing procedures.

Model-fit can be used as a means to investigate the appropriateness of a certain IRT model over others (e.g., UIRT vs. MIRT) for the data analyzed. Therefore, model-fit statistics, including Akaike's Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), root mean square error of approximation (RMSEA; Browne & Cudeck, 1993), and summed-score based item fit diagnostic (Orlando & Thissen, 2000), are provided in Table 2. Smaller values of the AIC, BIC, and RMSEA indicate better model fit. The fit statistics reported in Table 2 suggest that SS-MIRT fits better than UIRT for the new form. However, mixed results were observed for the old form. The AIC and BIC indices indicate a better fit of SS-MIRT than UIRT, while RMSEA and item fit statistics favors UIRT over SS-MIRT. Overall, both forms have the potential for being multidimensional.

Model fit can also be evaluated by inspecting the degree of consistency between the fitted marginal observed-score distribution and the actual observed frequency distribution. The fitted distributions using UIRT and SS-MIRT are depicted in Figure 1. The fitted distributions for both models seem to closely approximate the actual distribution for both the

new and old forms. One possible explanation for such a close alignment between two models is that the data are only moderately multidimensional, so the MIRT models might not perform much differently from the UIRT models.

Equating results for raw scores are displayed in Figure 2 using the identity equating as a baseline. All IRT-based procedures seem to produce very similar equating results, except at the higher end of the score range. On the contrary, the EQ procedure demonstrates a different pattern than the other equating procedures, especially in the score range of 10 – 15. This might possibly be due to the low – and at times zero – frequency at this score range. Note that since the IRT true scores cannot be identified for the observed scores of 9 and below (i.e., score points below the sum of guessing parameters), and no frequency was found for raw scores of 80 and above, the equating results were reported only for a raw score range of 10-79.

Figure 3 shows the equating results for the unrounded scale scores using the EQ procedure as a baseline. The reason that EQ served as a baseline here is that it was assumed not to be affected by multidimensionality, or at least does not explicitly address the unidimensionality assumption. An interpretation could be made only for the scale score range of 10-62. In Figure 3, the two dotted lines represent the Differences That Matter (DTM) criterion, which is conventionally used to determine if an acceptable level of equating results is achieved (Dorans & Feigenbaum, 1994). In general, similar patterns were found for scale score equating results, as were observed for the raw score results. The SMT equating procedure tends to produce the equating relationships similar to the UT and SMO procedures, although some large differences occur at the upper end of the score scale, where SMO slightly deviates from the DTM line.

The SMO and SMT procedures share similar features regarding the IRT models used to define equating relationship. They differ, however, in that one searches for the relationship between true scores on two forms and applies it to the relationship of observed scores whereas the other defines a model-based distribution and obtains the equating relationship through equipercentile equating. Therefore, the equating relationship between SMO and SMT procedures was compared and can be found in Figure 4. In this figure, the SMO procedure served as a baseline for the purpose of comparison. The two SS-MIRT procedures tend to yield very similar results over the entire score range with the largest difference around the score range of 55-63–even in this score range, the difference between two SS-MIRT equating procedures falls inside the DTM boundaries.

Often, the large differences between equating procedures are found to be more related to the psychometric model, not to the procedure type (Brossman & Lee, 2013). Thus, it would be meaningful to examine how SMT and UT perform on the same data. Figure 5 shows equated scale-score differences between the SMT and UT procedures. In this figure, the UT procedure served as a baseline only for comparative purposes. Overall, the SMT procedure yields a similar equating relationship to the UT procedure, although a larger difference was found at the end of the high score range, falling outside the DTM boundaries for scale scores of approximately 50 or above. Their close relationship might be attributed to the data structure, which showed only a moderate level of multidimensionality.

Table 3 presents the first four moments of equated scores for the studied equating procedures. For the equated raw scores and the unrounded scale scores, the moments for EQ seem the closest to the moments for the old form. Results are somewhat mixed for the rounded scale scores. In sum, the proposed SMT procedure seems to yield comparable results to the other methods.

A Simulation Study

Data Preparation

This simulation study was intended to evaluate the performance of the SMT equating procedure under various study conditions. The item parameters were estimated from the data used in real data analysis (i.e., AP English Language exams). The estimated item parameters served here as generating item parameters. Data were generated using the 3PL and GR models for MC and FR items, respectively. As with the real data analysis, four other equating procedures were also considered. In addition, when the bivariate frequency distribution was estimated, all three suggested approaches were applied and compared: (a) actual bivariate frequency distribution (SMT-ad), (b) bivariate log-linear smoothed frequency distribution (SMT-sm), and (c) estimated bivariate distribution based on the SS-MIRT models (SMT-md).

Composite scores were formed by simply summing the two section scores, which led to a possible score range of 0-81 and 0-82 for the old and new forms, respectively. Note that the simulation study did not include equating results for scale scores, because the real data analysis in the previous section suggested that the results for scale scores did not much differ from the results for raw scores.

Simulation Conditions

Correlation between MC and FR sections. Three levels of correlation between MC and FR section abilities were considered: .5, .8, and .95. A previous study conducted by Lee

and Brossman (2012) indicates that, based on a DTM criterion, equating results obtained from unidimensional or classical equating procedures would be reasonably acceptable with a latent-trait correlation of .8 or above under the random groups design. Thus, a correlation of .8 was included in this study as a benchmark at which MIRT models can be expected to perform better than UIRT models. Some may argue that data with a latent-trait correlation of .5 or lower is rarely seen in realistic settings; however, this could still occur in some language tests in which FR items are designed to address distinct abilities such as speaking or writing, so the impact of such a substantial multidimensionality is worth investigating. On the contrary, when a correlation is .95 or above, meaning that the data can be regarded as approximately unidimensional, format effects become less influential.

Sample size. Two levels of sample size were included in this study: 1,000 and 5,000. Previous research pointed out that having sample size of 5,000 or more leads to reasonably accurate equating results (Hanson & Beguin, 2002; Kirkpatrick, 2005). Thus, more precise equating results were anticipated to be associated with the larger sample size. With sample size of 1,000, the extent to which moderate or small sample size affects equating precision could be assessed.

Simulation Procedure

The specific simulation process is described below:

(1) Randomly draw pairs of theta values $(\theta_M,\ \theta_F)$ for a group of examinees from a given bivariate normal distribution, $BN(0,0,1,1,\rho_{\theta_M\theta_F})$.
(2) Generate item responses for each examinee for each MC item using MC item parameters for the new form and the true MC theta value (θ_M) generated in (1).
(3) Generate item responses for each examinee for each FR item using FR item parameters for the new form and the true FR theta value (θ_F) generated in (1).
(4) Repeat step (2) and step (3) using the old form item parameters.
(5) Conduct the seven equating procedures, including the three SMT methods, using the generated data to find the estimated equating relationships for each procedure.
(6) Repeat the above steps 100 times.

For the SMT procedures, additional steps were required to find a bivariate score distribution. For the SMT-ad procedure, the actual bivariate score distribution was found through a contingency table for the MC and FR scores. For the SMT-sm procedure, the actual bivariate score distribution was smoothed with a smoothing parameter of 6-6-1 using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). Last, the fitted bivariate score

distribution for the model-based method (i.e., IRT-fitted) was found using code written in R (R Core Team, 2014). Note that the correlation between two section abilities was estimated for each generated dataset at the time when item parameters were calibrated using *flexMIRT* (Cai, 2017). Then, the estimated correlation for the new form was used to find the bivariate score distribution, $f(x_M, x_F)$, in Equation 5.

Criterion Equating Relationships

The criterion equating relationships were established based on large-sample single-group equipercentile equating for which both forms are assumed to be given to the same group. The large-sample single-group equating criterion suggested by Kim and Lee (2016) has several advantages over other possible criteria in that it is free from sampling error due to use of a large sample and it removes any potential equating error that might occur when using two groups instead of one. The specific steps are as follow:

1) Draw a large sample (N=1,000,000) from the bivariate normal distribution for each level of correlation, $(\theta_M, \theta_F) \sim BN(0,\ 0,\ 1,\ 1,\ \rho_{\theta_M\theta_F})$. Note that a group of large samples is defined separately for each level of correlation between MC and FR abilities.
2) Generate item responses for each examinee for both the old and new forms. Thus, each examinee has scores for both forms.
3) Find an equating relationship using the traditional equipercentile equating.
4) Repeat the above steps for each of the three different levels of correlation.

It is important to remember that the actual frequency distribution for the new form served as weights in computing summary statistics introduced in the subsequent section.

Evaluation Criteria

The estimated equating relationships from 100 replications for each equating procedure were compared to the criterion equating relationship. To quantify and compare the performance of each equating procedure, squared bias (SB), variance (VAR), and mean squared error (MSE) were computed. For each score x,

$$SB(x) = \left[\left(\frac{1}{100}\sum_{r=1}^{100}\hat{e}_{xr}\right) - e_x\right]^2, \tag{8}$$

$$VAR(x) = \frac{1}{100}\sum_{r=1}^{100}\left[\hat{e}_{xr} - \left(\frac{1}{100}\sum_{r=1}^{100}\hat{e}_{xr}\right)\right]^2, \tag{9}$$

$$MSE(x) = SB(x) + VAR(x), \quad (10)$$

where e_x is the criterion equated score at score *x* and $\hat{e}_{xr}$ is an estimated equated score at score *x* on replication r. Finally, the overall statistics (SB, VAR, and MSE) were obtained as a weighted sum of the above statistics across all score points, with each score point on the new form weighted by its corresponding frequency.

In addition to the overall statistics, conditional statistics were considered to explore the precision of each equating procedure across different score points. For the conditional results, the squared terms in Equations 8 and 9 were removed to facilitate examination of the direction of bias. As a result, root mean squared error (RMSE), signed bias, and standard error (SE) were computed, without performing the squaring operation in Equations 8 through 9.

Results of the Simulation Study

The primary purpose of conducting the simulation study was to evaluate the performance of SMT by comparing it to other existing equating procedures. This section focuses on the comparison between these equating procedures. The results based on the overall statistics are presented first, followed by the results for the conditional statistics. The last part of this section is devoted to comparing the three proposed SMT methods.

Overall statistics. Results are provided first for the sample size of 1,000 and then for the sample size of 5,000. However, the general pattern seemed to be consistent across the sample size conditions, so the primary focus of the discussion is on the smaller sample size.

Three overall statistics for the sample size of 1,000 can be seen in Table 4. The columns in Table 4 represent seven equating procedures including the three SMT methods. The equating results for the three levels of correlation are also reported. In terms of VAR, SMO tends to result in the smallest values, closely followed by the three SMT methods. It is worth mentioning that smaller VAR is generally found with SMT-md than UO, with the exception of the largest correlation condition (i.e., $\rho = .95$). This finding suggests that SMT is preferable over UO, even when a moderate level of multidimensionality occurs. This finding is particularly surprising because the observed-score equating procedure has been recognized as having less variability than the true-score equating procedure, given that UO has been reported to produce smaller standard errors than UT in previous studies (Cho, 2008; Hagge et al., 2011; Tsai, Hanson, Kolen, & Forsyth, 2001). Therefore, it might be reasonable to expect more consistent equating relationships for SMT than for UO or UT when the unidimensionality assumption cannot be retained.

According to SB, both SMO and UO outperform the three SMT methods. One possible explanation of this tendency is the fact that the criterion equating relationship was established using the equipercentile equating procedure based on observed scores generated by the SS-MIRT model. Consequently, study results might be biased favorably towards the (M)IRT observed-score procedures (i.e., UO or SMO) or EQ. In spite of its disadvantages towards true-score procedures, the SMT procedures, especially SMT-md, tend to provide SB values that are almost comparable to the UO procedure. One notable finding is that the SB for UT becomes substantially larger as the correlation decreases, while the SB for SMT-md remains relatively consistent across the correlation conditions. As a result, when compared to UT, SMT-md seems to reduce SB substantially, even for the approximately unidimensional data (i.e., $\rho = .95$).

In general, the MSE results indicate that the two (M)IRT observed-score procedures (i.e., UO and SMO) show the smallest error, followed by the three SMT procedures, next by UT, and last by EQ. Under a small or moderate correlation (i.e., $\rho = .5$ or .8), SMO shows the smallest error among the seven procedures. When the data are approximately unidimensional (i.e., $\rho = .95$), UO outperforms the other procedures, which is in line with an observation by Lee and Brossman (2012). Also, it is worth mentioning that a consistent pattern was found for the relationship between the SMT and UT equating procedures. That is, SMT-md always provides more accurate equating results than UT. This relationship holds for any statistics used, any correlation levels, and any sample size condition. In addition, it appears obvious that as a correlation becomes smaller, the difference between the two procedures tends to become bigger.

The overall findings in equating results for the sample size of 5,000, as can be found in Table 5, are generally very similar to the findings for the sample size of 1,000, with only a few exceptions. First, as anticipated, the use of larger sample size leads to reduction in variance. As a result, the SB contributes more to the MSE relative to the VAR, which makes the patterns of the MSE mirror those of the SB. Second, increasing sample size results in smaller differences between equating procedures, especially in the VAR values. A differential equating precision in relation to sample size implies that when sample size is not sufficiently large, deciding which equating procedure should be chosen should be made with more caution. In terms of SB, however, the general tendency seems to remain constant, except for significantly smaller SB for EQ with a correlation of .5.

Conditional statistics. In this section, equating results are presented only for the sample size of 1,000, since similar results were observed for the sample size of 5,000. For simplicity, the accompanying figures present only the SMT-md method. The main reason for selecting the SMT-md was that, after probing with the overall statistics, it was the most accurate of the three SMT methods. A comparison of the three SMT methods will be provided in detail in the following section.

Conditional SE for five equating procedures based on sample size of 1,000 can be seen in Figure 6. Note that the scale on the vertical axis is truncated to facilitate meaningful visual comparison of differences among equating procedures. Consequently, equating results that deviate from this range could not be captured in the figures, although only the results for EQ at the lower end of the score range fell outside this range. In general, EQ introduces a significant amount of SE at the extreme ends of the score distribution. This particular pattern remains constant across all correlation conditions. It seems apparent that the SE trends for three equating procedures including UO, SMO, and SMT-md are unaffected by correlation levels. On the other hand, the SE pattern for UT varies with the level of correlation. Specifically, as a correlation decreases, a larger SE tends to be observed for UT at the extreme score ranges.

The signed bias is displayed in Figure 7. In this figure, the horizontal black solid line serves as a zero line (i.e., no bias). In general, all the lines for the five equating procedures are below the zero line across all correlation levels for most of the score range, indicating a negative bias across the score scale; an exception occurs in the 20-30 range or the upper score range, where only a few examinees scored. One notable finding is that the behavior of UT is fairly different from that of the other procedures. That is, UT introduces a large amount of negative bias at the upper and lower score ranges. This becomes more noticeable as the correlation gets smaller. UT, unlike the other procedures, has slightly positive bias near the middle of the scale. This unique pattern remains consistent across all levels of correlation. For SMT-md, the general tendency seems similar to the other procedures except UT. However, there is a slightly larger negative bias at the very lower score range relative to the others, especially when the correlation is very low (i.e., $\rho = .5$).

Figure 8 shows the conditional RMSE for the five equating procedures for three levels of correlation. Overall, due to the larger values of SE in magnitude, the SE contributes to the overall error to a greater extent than the bias does. Consequently, the plots for the RMSE closely resemble the plots for the SE. As in Figure 6, there is a high degree of consistency between most of the procedures, except EQ and UT. A larger error tends to be associated

with the UT procedure near the upper score end as the correlation decreases, while the pattern of EQ does not substantially vary depending upon the level of correlation.

Comparison of SMT methods. The comparison of the three SMT methods are made through a visual inspection with Figure 9. Since the general findings were fairly stable across all study conditions, only the conditional results for $\rho = .5$ based on sample size of 1,000 are plotted. The top panel in the figure exhibits the SE results, the middle panel displays the bias results, and the bottom panel reveals the RMSE results. According to the SE statistics, smaller values are consistently observed for SMT-sm and SMT-md relative to SMT-ad. This pattern becomes more remarkable at the upper score points. There is a high degree of consistency between two methods, SMT-sm and SMT-md.

All three SMT methods tend to produce negative bias. However, the IRT-fitted method in general demonstrates less deviation from the zero line than the other two methods, especially at the upper end of a distribution. For the middle part of the distribution (i.e., a range of 40-60), the IRT-fitted method introduces slightly larger bias than the other two, but not to a great extent.

As shown in the bottom part of Figure 9, the SMT-md method consistently provides the smallest overall error across the entire scale with very few exceptions, suggesting that the IRT-fitted method outperforms the two other proposed methods. The biggest difference between SMT-md and the other methods is found at the upper end of the scale, mainly due to its smaller bias in this range.

Conclusions and Discussion

When multiple forms of a test are administered, score comparability over the forms becomes an essential part of the scoring and reporting process. Often, tests administered multiple times with several variations of test forms are high-stakes such as college entrance exams. Thus, it is imperative for those tests to maintain a satisfactory level of precision in equating. However, the accuracy of equating cannot be guaranteed if the data structure does not satisfy the statistical assumptions of an equating procedure used (Kolen & Brennan, 2014). The SS-MIRT model often provides an adequate fit to multidimensional data such as a mixed-format test; as such, this paper uses the SS-MIRT framework to develop a true-score equating procedure that can be used for mixed-format data.

Two separate analyses were conducted using real data and simulated data to evaluate the feasibility and accuracy of the proposed equating procedure. Five equating procedures were compared in this study: (a) equipercentile equating with a log-linear presmoothing method, (b) UIRT true-score equating, (c) UIRT observed-score equating, (d) simple-

structure true-score equating, and (e) simple-structure observed-score equating. In addition, regarding the SMT procedure, three methods to determine the bivariate score distribution for two sections were applied and compared: using (1) an actual bivariate score distribution, (2) log-linear smoothed bivariate distribution, and (3) IRT-fitted bivariate distribution.

Major findings based on the results from real data analysis can be summarized as follows. First, the proposed SMT procedure tends to provide very similar equating results to other IRT-based procedures such as SMO, UT, and UO. Second, a relatively large difference between UT and SMT occurs at the higher end of the score distribution. Unfortunately, previous studies have not yet reached consensus regarding the formal relationships between equipercentile, UIRT, and MIRT equating procedures. Lee and Brossman (2012) observed substantial differences between the UO and SMO equating procedures and closer conformity between SMO and EQ. Their observation was made based on data with a disattenuated correlation of .91 for the old form and .94 for the new form, which are even more unidimensional than the data used in this study. This finding lies in contrast to later observations by Peterson and Lee (2014), who compared the performance of the full MIRT observed-score equating procedure with EQ, UO, and Bifactor observed-score equating procedures. They found that the EQ and UIRT equating procedures provided equating results that were the most different from the identity criterion. Further examination needs to be done to arrive at a clear explanation regarding these mixed findings.

One of the most important findings from the simulation study is that SMT consistently outperforms the UT equating procedure no matter which overall statistic is used. In addition, the study results support the previous finding that SMO introduces smaller error than UO when the proficiency correlation is very low (i.e., .5) (Lee & Brossman, 2012). RMSE was lowest for the UO method when the data were unidimensional ($\rho = .95$) and for the SMO method when the data were multidimensional ($\rho = .5$ and .8). This also suggests less error with observed score than with true score equating, which might be partly due to the equating criterion used in this study.

Another interesting finding is that decreasing ability correlation results in remarkably larger bias for the UT procedure, while bias for the SMT procedure is relatively robust to smaller correlation. In terms of the performance of the three proposed SMT procedures, the IRT-fitted method might be preferable to other methods because it presents a smaller error across all study conditions.

Some limitations of this study should be acknowledged. One of such limitations is that the criterion equating relationships established for the simulation study were based on the

SS-MIRT framework. In defining the criterion equating relationships, the data were generated using the SS-MIRT model and traditional equipercentile equating method was applied, which might lead to results that favor the SS-MIRT equating procedures, especially SMO. The results based on other simulation criteria might not lead to the same conclusion as made in this study. Another limitation stems from the fact that both the real data analysis and the simulation study were conducted in the context of random groups design. Indeed, this is neither a limitation nor a drawback of the proposed equating procedure itself. Since the main focus of this paper was to introduce the SS-MIRT true-score equating procedure, our goal was to be as parsimonious as possible in terms of other study conditions which could exert unintentional impacts on results, and to concentrate more on the performance of the proposed equating method itself. As emphasized earlier, SMT equating can be conducted in conjunction with other designs such as the CINEG design after scale linking procedure. Future research on SMT equating can expand the research scope by incorporating various study designs.

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* (pp. 397-479). Reading, MA: Addison-Wesley.

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. (Available from the web site: http://www.uiowa.edu/~casma).

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

Browne, M. W., & Crudeck, R. (1993). Alternative ways of assessing model fit. *Sage focus editions, 154*, 136-136.

Cai, L. (2017). *flexMIRT*. [Computer Program]. Chapel Hill, NC: Vector Psychological Group, LLC.

Cheng, Y.-Y., Wang, W-C., & Ho, Y.-H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests. A statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement, 69*, 369-388.

Cho, Y. (2008). *Comparison of bootstrap standard errors of equating using IRT and equipercentile methods with polytomously-scored items under the common-item nonequivalent-groups design* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Davey, T., Oshima, T., & Lee, T. (1996). Linking multidimensional item calibration. *Applied Psychological Measurement, 20*, 405-416.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement, 32*, 355-370.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of MCMC in test scoring. *Journal of Educational and Behavioral Statistics, 30*, 295-311.

Dorans, N. J. (2000). *Distinctions among classes of linkages* (College Board Research Note RN-11). New York: College Board.

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* (Research Memorandum No. RM-94-10). Princeton, NJ: Educational Testing Service.

Hagge, S. L., Liu, C., He, Y. Powers, S. J., Wang, W., & Kolen, M. J. (2011). A comparison of IRT and traditional equipercentile methods in mixed-format equating. 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). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note.

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.

Kim, S. Y., & Lee, W. (2016). Composition of common items for equating with mixed-format tests. In M. J. Kolen & Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 4).* (CASMA Monograph No. 2.4). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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.

Lee, G., & Lee, W. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29,* 224-241.

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.

Li, Y., & Lissitz, R. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24,* 115-138.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29,* 3-25.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 452-461.

Min, K. (2007). Evaluation of linking methods for multidimensional IRT calibrations. *Asia Pacific Education Review, 8,* 41-55.

Moses, T., & Holland, P. W. (2010). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47,* 76-91.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.

Oshima, T., Davey, T., & Lee, K. (2009). Multidimensional linking: four practical approaches. *Journal of Educational Measurement, 37,* 357-373.

Peterson, J., & Lee, W. (2014). Multidimensional item response theory observed score 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 3).* (CASMA Monograph No. 2.3). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Thompson, T., Nering, M., & Davey, T. (1997). *Multidimensional IRT scale linking*. Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Tsai, T.-H., Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education, 14,* 17-30.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28,* 295-316.

Wang, W.-C, Chen, P.-H, & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Table 1

Descriptive Statistics

Statistics	New Form	Old Form
Raw Score Scale	0-82	0-81
# of MC Items	55	54
# of FR Items (maximum score)	3 (9,9,9)	3 (9,9,9)
Mean	47.090	49.025
SD	12.502	12.448
Skewness	-.248	-.470
Kurtosis	-.307	-.225
$\hat{\rho}_{\theta_{MC}\theta_{FR}}$	.82	.80
N	6,000	6,000

Table 2

Fit Statistics

Model	AIC	BIC	RMSE	$S-\chi^2$ ($\alpha = .05$)	
				# of Misfit	Misfitted Items
New Form					
UIRT	437180.70	438487.11	.10	6	M8,M16,M28,M32,M48,F2
SS-MIRT	436534.54	437847.64	.10	5	M16,M28,M32,M48,F2
Old Form					
UIRT	408460.61	409746.92	.09	14	M5,M7,M10,M12,M28,M29, M40,M47,M48,M50,M51, M52,M53,M54
SS-MIRT	404053.24	405721.41	.10	15	M3,M5,M7,M10,M12,M28, M29,M32,M40,M47,M48, M50,M51,M54,F2

Table 3

Moments for Equated Scores

	Mean	S.D.	Skew	Kurt
Raw Scores				
Old form	49.025	12.448	-.470	2.773
	New Form Equated to Old Form Scale			
SMT-md	48.963	12.471	-.465	2.779
SMO	49.022	12.460	-.461	2.779
EQ	49.027	12.440	-.468	2.770
UT	48.947	12.434	-.492	2.767
UO	49.003	12.419	-.467	2.776
Unrounded Scale Score				
Old form	33.716	8.333	-.010	3.012
	New Form Equated to Old Form Scale			
SMT-md	33.681	8.363	.020	3.032
SMO	33.723	8.370	.030	3.039
EQ	33.719	8.326	-.003	2.989
UT	33.635	8.300	.170	2.744
UO	33.694	8.343	.220	2.814
Rounded Scale Score				
Old form	33.685	8.305	-.018	3.045
	New Form Equated to Old Form Scale			
SMT-md	33.653	8.356	.019	3.030
SMO	33.674	8.335	.034	3.048
EQ	33.717	8.349	-.001	3.009
UT	33.627	8.299	.177	2.761
UO	33.662	8.340	.230	2.805

Table 4

Summary Statistics for Simulation Results (N=1,000)

Exam	EQ	UT	UO	SMT-ad	SMT-sm	SMT-md	SMO
$\boldsymbol{\rho = .5}$							
VAR	.52864	.39441	.35796	.38283	.35891	.35380	.34347
SB	.02418	.17776	.01908	.04098	.03980	.03041	.01359
MSE	.55282	.57217	.37704	.42381	.39871	.38421	.35706
$\boldsymbol{\rho = .8}$							
VAR	.68964	.55574	.52938	.53155	.51180	.49780	.50962
SB	.00873	.09910	.01231	.03048	.03003	.02280	.00867
MSE	.69837	.65484	.54170	.56203	.54184	.52060	.51829
$\boldsymbol{\rho = .95}$							
VAR	.70775	.54171	.51659	.56115	.54308	.52896	.52633
SB	.00506	.06378	.00930	.02000	.01991	.00933	.01012
MSE	.71282	.60549	.52589	.58115	.56299	.53829	.53645

Note. Underlines indicate the smallest equating error among the equating procedures.

Table 5

Summary Statistics for Simulation Results (N=5,000)

Exam		EQ	UT	UO	SMT-ad	SMT-sm	SMT-md	SMO
$\boldsymbol{\rho = .5}$								
	VAR	.10253	.09079	.07750	.07753	.07308	.07145	.06732
	SB	.00537	.16149	.00986	.02143	.02063	.01462	.00520
	MSE	.10790	.25228	.08736	.09896	.09370	.08606	.07253
$\boldsymbol{\rho = .8}$								
	VAR	.11562	.09620	.08822	.09270	.08882	.08580	.08434
	SB	.00278	.09106	.00365	.01837	.01791	.01163	.00193
	MSE	.11839	.18726	.09187	.11108	.10674	.09744	.08628
$\boldsymbol{\rho = .95}$								
	VAR	.11215	.08901	.08239	.09428	.09176	.08842	.08633
	SB	.00611	.06102	.00561	.01446	.01515	.00526	.00555
	MSE	.11826	.15003	.08800	.10873	.10691	.09368	.09188

Note. Underlines indicate the smallest equating error among the equating procedures.

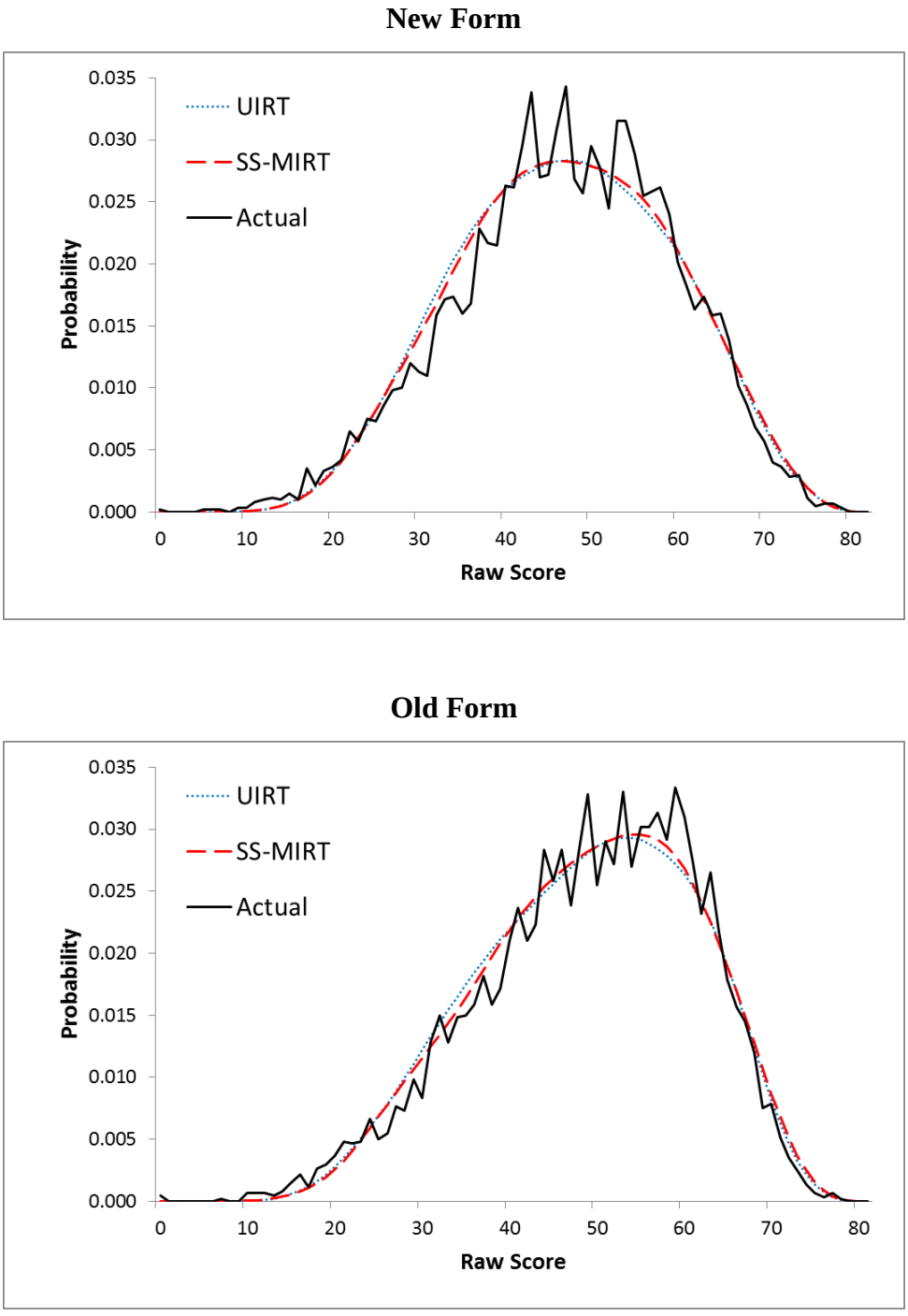


Figure 1. Observed and fitted distributions for old and new form examinees.

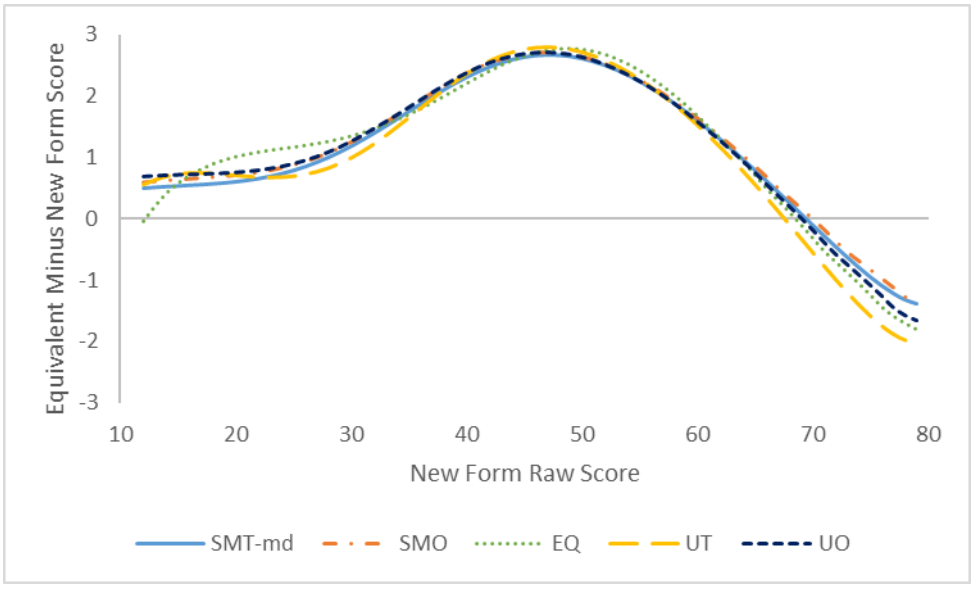


Figure 2. Raw-to-raw score equivalents for five equating procedures.

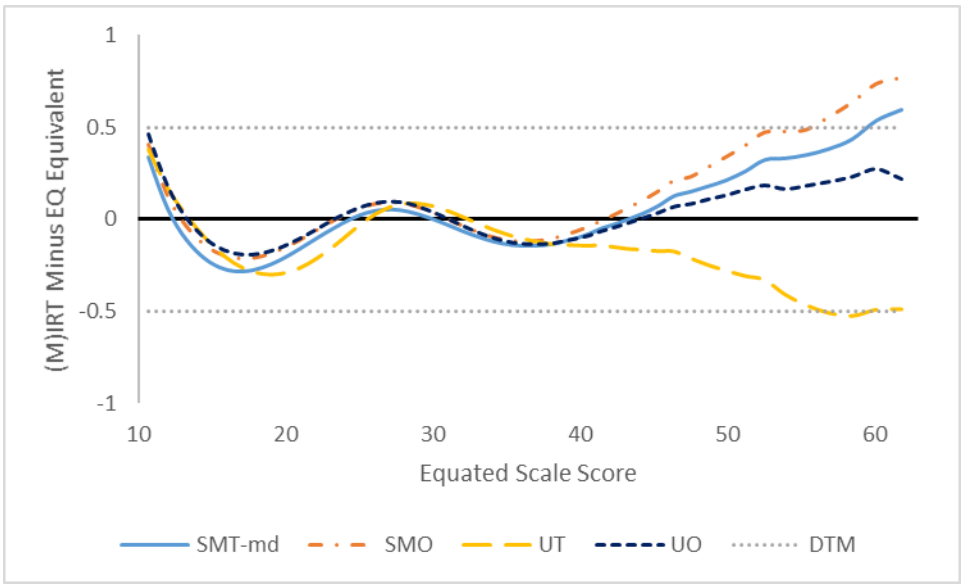


Figure 3. Differences between (M)IRT and equipercentile equated unrounded scale scores.

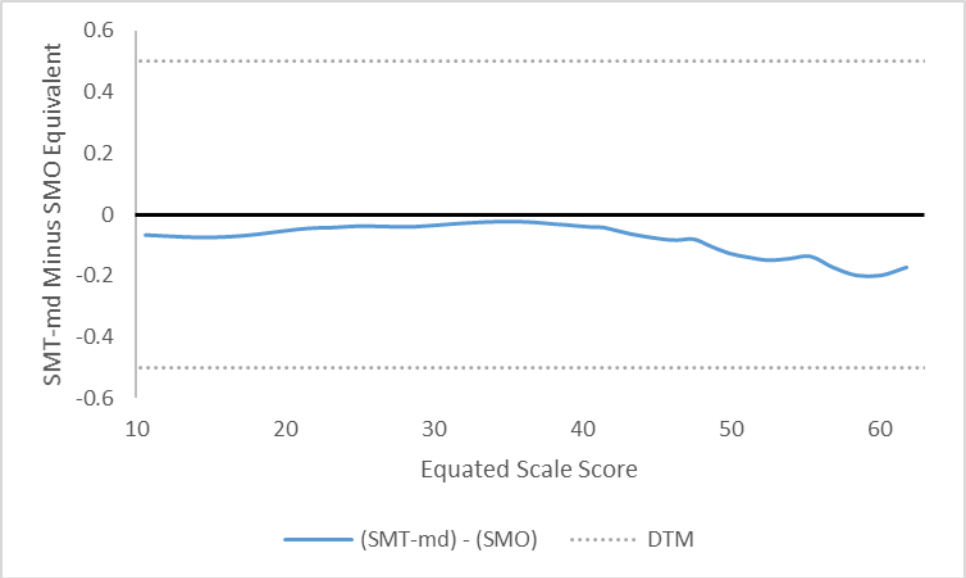


Figure 4. Differences between SMT and SMO equated unrounded scale scores.

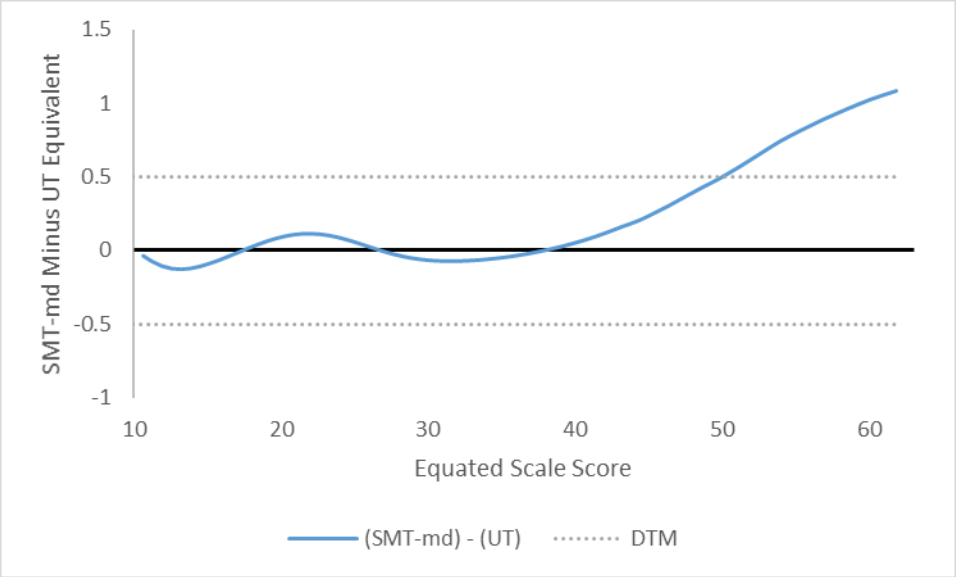


Figure 5. Differences between SMT and UT equated unrounded scale scores.

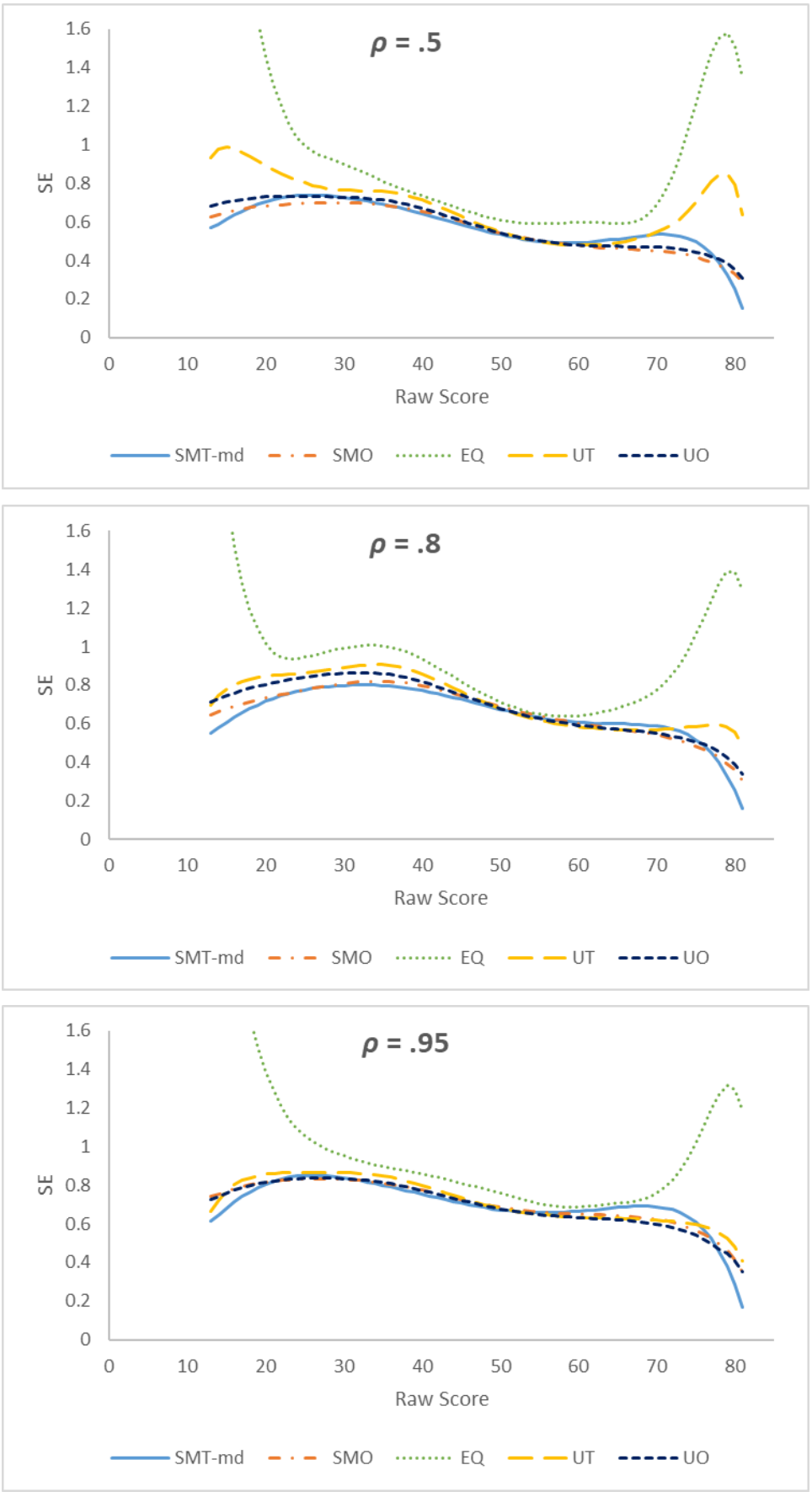
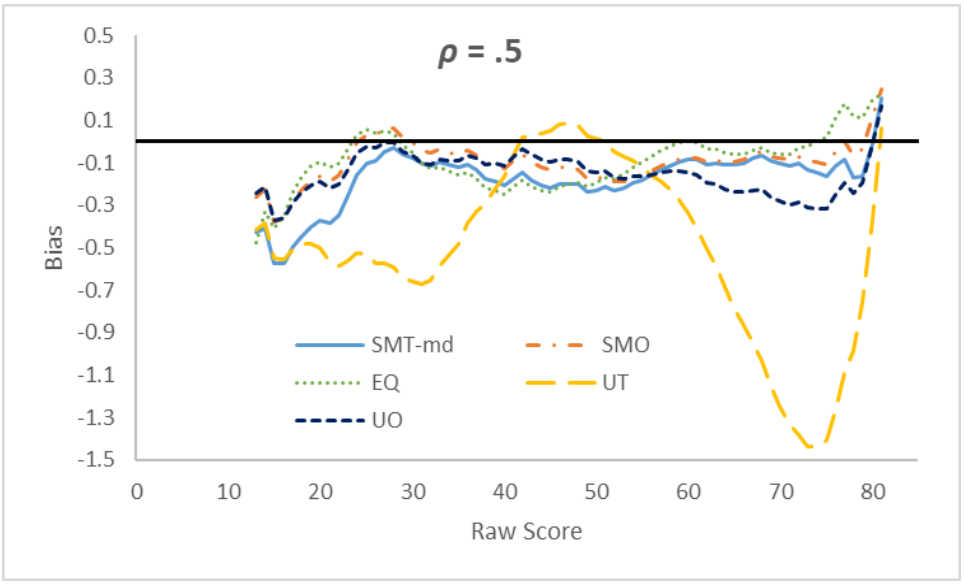
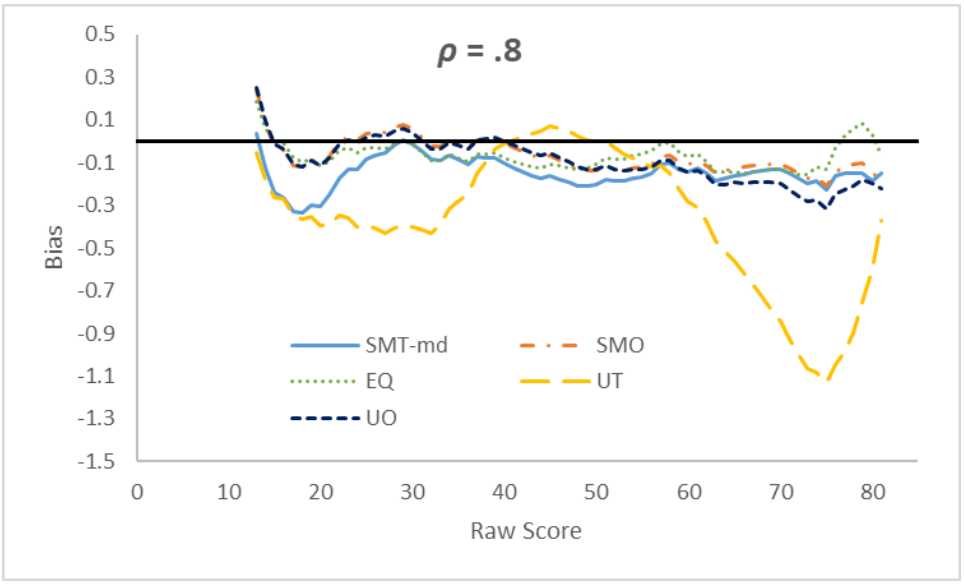


Figure 6. Conditional SE for N =1,000.





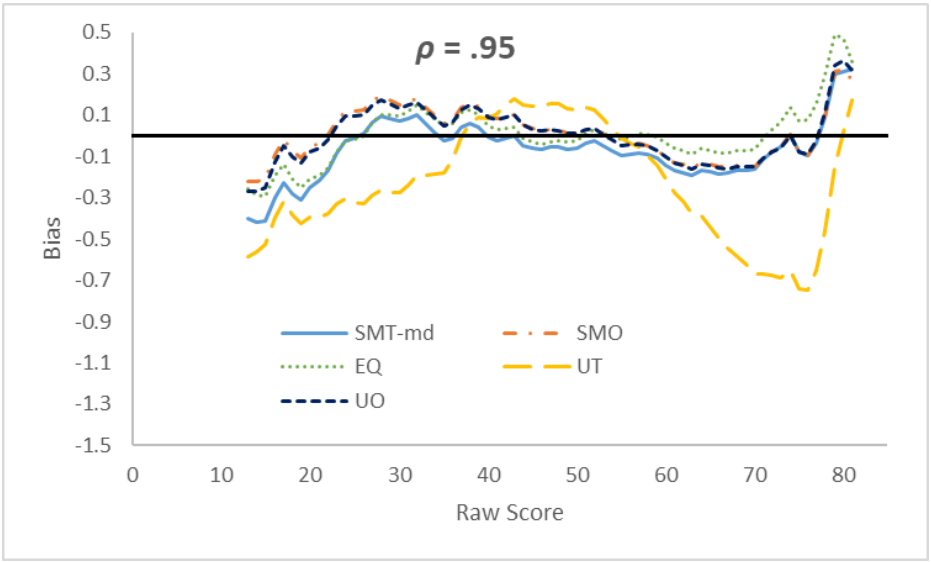


Figure 7. Conditional bias for N =1,000.

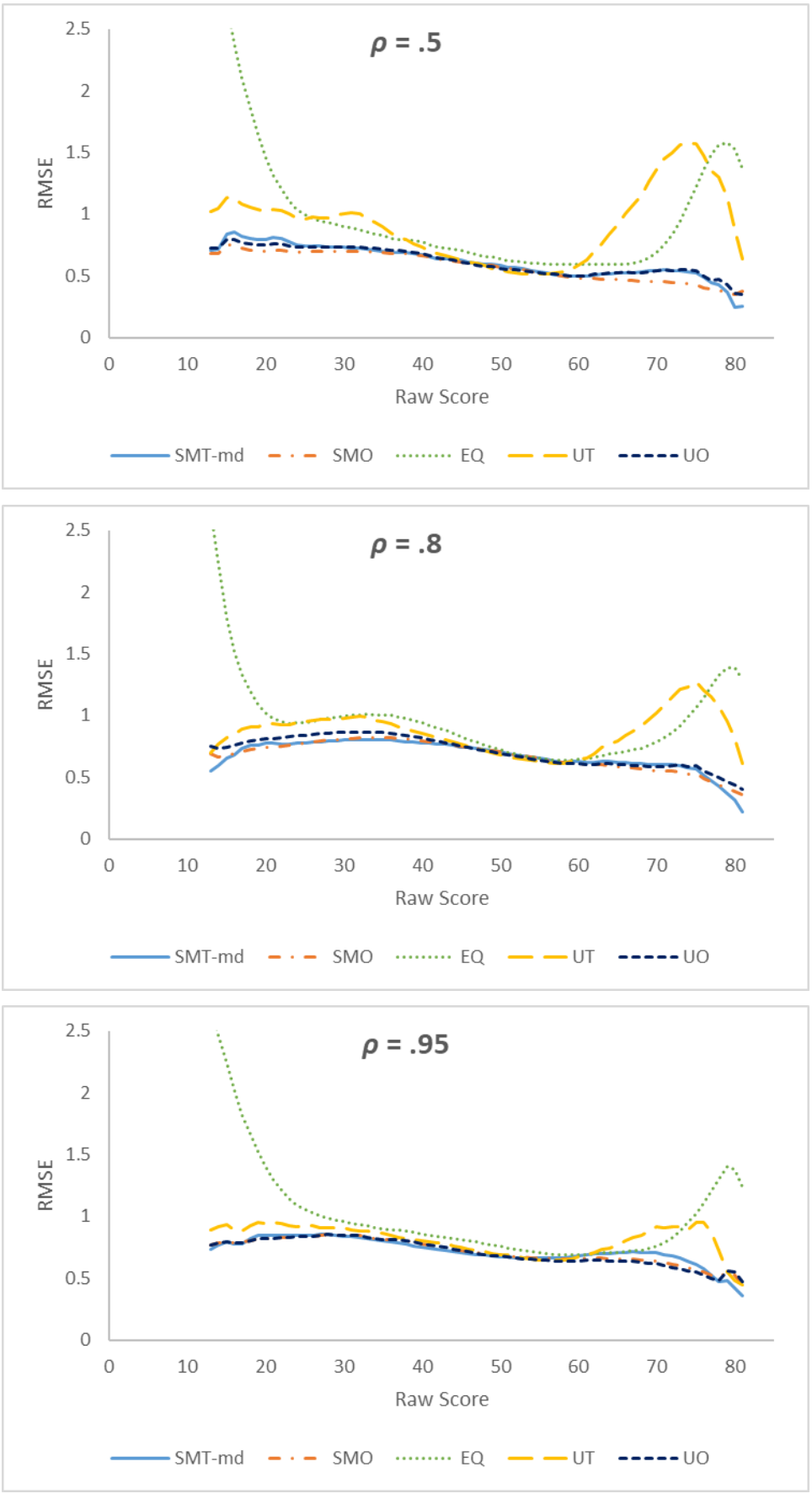


Figure 8. Conditional RMSE for N =1,000.

Figure 9. Conditional statistics for three SMT methods for $\rho = .5$ and N =1,000.

Chapter 6: Chained Beta True Score Equating for the Common-Item Nonequivalent Groups Design

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

Abstract

The main purpose of this study was to investigate the accuracy of chained beta true score equating methods and compare it to the accuracy of traditional and IRT methods under various conditions for the common item non-equivalent groups design. Real data and simulated data were used to study the effect of sample size and group ability difference on six equating methods, including frequency estimation, chained equipercentile, IRT true score and observed score, and Beta2 and Beta4 true score methods. The conditional and weighted absolute bias, standard error of equating, and root mean square error were used to evaluate the equating results generated by simulated data. The results showed that the Beta2 method generated a smooth equating relationship, produced more accurate equating results compared to the frequency estimation and chained equipercentile methods, and was not affected by group differences. The Beta4 method generated very different equating relationships compared to the other methods, and yielded the least accurate equating results.

Chained Beta True Score Equating for the Common-Item Nonequivalent Groups Design

Equating plays an important role in test scoring when multiple forms of a test exist. When equating is performed successfully, examinees are expected to earn the same equated score regardless of the test form administered, and examinees who earn the same equated score on multiple forms are considered to be at the same achievement level (Kolen & Brennan, 2014). To equate test scores, two issues must be considered carefully: the approach for collecting the data and the appropriate statistical estimation method for analyzing it. In the equating literature, the approach for collecting data is referred to as the "equating design," and the statistical estimation method for analyzing the data is referred to as the "equating method." The focus of this chapter is on the common-item nonequivalent groups (CINEG) design, in which only one form is needed in one administration. Though this design requires a more complicated statistical analysis compared to other designs, it is widely used in practice because of its flexibility in administration.

A number of methods have been developed for equating, which can be classified by theoretical basis, including traditional equating, item response theory (IRT) equating, and beta equating. Traditional equating (Angoff, 1971; Kolen & Brennan, 2014) is accomplished by setting certain characteristics of the two score distributions equal for a particular group of examinees. In general, traditional equating falls into three categories: mean, linear, and equipercentile equating. The interest of this chapter is in equipercentile equating because it is more generally used in practice and provides greater similarity between distributions of equated scores than mean and linear equating. Three equipercentile equating methods used with the CINEG design were considered, including the frequency estimation equipercentile (FE) and chained equipercentile (CE) equating methods. IRT equating, as indicated by its name, is built upon IRT, which offers mathematical models to relate the probability of an examinee's correctly answering an item to his or her underlying ability of interest and characteristics of the item (Lord, 1980).

Beta equating refers to equating methods developed from strong true score theory (STST). STST describes a relationship between observed scores and true scores by assuming the distribution of true scores is either a four-parameter or two-parameter beta, and the distribution of observed scores given a true score is a binomial or compound binomial. (Hanson, 1991; Keats

& Lord, 1962; Lord, 1965; 1969). Accordingly, the distribution of observed scores is the combination of these two assumed distributions and can take on a variety of forms. Beta equating includes beta observed score and true score equating. Beta observed score equating, also referred to in the literature as a smoothing method (Hanson, Zeng, & Colton, 1994; Kolen & Brennan, 2014), fits the two observed score distributions with a specific STST model first and then performs equipercentile equating with the fitted distributions. Chained beta true score equating was first proposed by Lord (1965) who suggested that the true scores of two test forms designed to measure the same proficiency would have a functional relation for any group of examinees. That is, the true scores of two alternate forms can be related to produce an equating relationship. Thus, chained beta true score equating is conducted by first estimating true score distributions using a specific STST model and then performing equipercentile equating with estimated true score distributions.

In the psychometric literature, very few studies have discussed the application of beta equating. Kim, Brennan, and Kolen (2005) compared beta equating with IRT equating for the random groups (RG) design using equity properties as criteria. Wang, Lee, Brennan, and Jing (in press) introduced a chained true score equipercentile method to assess equity properties for the common-item nonequivalent groups (CINEG) design under the framework of STST. This method can be adopted as a chained beta true score equating method for the CINEG design and yet, no study has researched this equating method. This gap in the beta equating literature motivates this study. Thus, this study aims to gain a better understanding of chained beta true sore equating for the CINEG design by 1) investigating the accuracy of beta equating results, and 2) comparing beta equating methods with traditional and IRT methods under various conditions.

Method

This study uses real and simulated data to investigate how the equated scores obtained from the beta equating methods differ under various conditions, how they compare with the equated scores obtained from traditional and IRT equating methods, and how they differ from criterion equating. The analyses were conducted using open source *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009) with some new code added.

Data

Data used in this study were from Advanced Placement (AP) Chemistry and Chemistry examinations, denoted as Tests 1 and 2. The tests employed two alternate forms (Forms X and

Y) sharing a certain number of items in common (V). The two forms were administrated to two different groups consisting of more than 100,000 examinees.

Test 1 data were used for real data analysis. The descriptive statistics for the two forms are shown in Table 1. The observed score distributions of the two forms are presented in Figure 1. Test 1 contains 60 multiple-choice items, with Forms X and Y sharing 12 items in common. The common items descriptive statistics and frequency distribution suggest higher ability in the Form X group than in the Form Y group.

Test 2 is composed of 50 multiple-choice items, with Forms X and Y sharing 12 items in common. 3PL IRT calibration was separately conducted for each form using flexMIRT (Cai, 2013) to estimate item parameters and the item parameters were put on the same scale through the common items using the Stocking-Lord method. These item parameter estimates were treated as generating (true) item parameters that were used to simulate item responses for simulation analysis. To illustrate item characteristics for each form, the means and standard deviations of the *a*-, *b*-, and *c*-parameters are presented in Table 2.

Chained Beta True Score Equating for the CINEG Design

Chained beta true score equating methods use proportion-correct true scores to find a true score relationship between Form X and Form Y through the proportion-correct true scores on a common-item set. It is assumed that examinees from Population 1 take Form X and common items V, and examinees from Population 2 take Form Y and common items V. τ_X is the true scores on Form X, and l_{τ_X} and u_{τ_X} are the lower and upper bound of τ_X. Similarly, τ_{V_1} is the true scores on V for population 1, and $l_{\tau_{V_1}}$ and $u_{\tau_{V_1}}$ are the lower and upper bound of τ_{V_1}; τ_Y is the true scores on Form Y, and l_{τ_Y} and u_{τ_Y} are the lower and upper bound of τ_Y; τ_{V_2} is the true scores on V for population 2, and $l_{\tau_{V_2}}$ and $u_{\tau_{V_2}}$ the lower and upper bound of τ_{V_2}. The steps are as follows:

1) Estimate the true score distributions of Form X, $f(\tau_X)$, and the common-item set, $h_1(\tau_{V_1})$, based on examinees from Population 1 using the assumed STST model.
2) Estimate the true score distributions of Form Y, $g(\tau_Y)$, and the common-item set, $h_2(\tau_{V_2})$, based on examinees from Population 2 using the assumed STST model.
3) Obtain the Group 1 common-item set equivalent proportion-correct true scores, τ_{V_1} ($0 \leq l_{\tau_{V_1}} < \tau_{V_1} \leq u_{\tau_{V_1}} \leq 1$), by converting the proportion-correct true scores on Form

X, τ_X $(0 \leq l_{\tau_X} < \tau_X \leq u_{\tau_X} \leq 1)$, using the equipercentile method. Refer to the resulting function as $e_{\tau_{v_1}}(\tau_X)$, so that

$$\int_{l_{\tau_X}}^{\tau_X} f(\tau_X) d\tau_X = \int_{l_{\tau_{V_1}}}^{e_{\tau_{v_1}}(\tau_X) = \tau_{V_1}} h_1(\tau_{V_1}) d\tau_{V_1}. \quad (1)$$

4) Identify the Form Y equivalent proportion-correct true scores, τ_Y $(0 \leq l_{\tau_Y} < \tau_Y \leq u_{\tau_Y} \leq 1)$, by converting the Group 2 common-item set equivalent proportion-correct true scores, τ_{V_2} $(0 \leq l_{\tau_{V_2}} < \tau_{V_2} \leq u_{\tau_{V_2}} \leq 1)$, using the equipercentile method. Refer to the resulting function as $e_{\tau_{Y_2}}(\tau_{V_2})$, so that

$$\int_{l_{\tau_{V_2}}}^{\tau_{V_2}} h_2(\tau_{V_2}) d\tau_{V_2} = \int_{\tau_Y}^{e_{\tau_{Y_2}}(\tau_{V_2}) = \tau_Y} g(\tau_Y) d\tau_Y. \quad (2)$$

5) Link the two conversion functions together to produce a conversion of Form X proportion-correct true scores to Form Y proportion-correct true scores, $e_{\tau_Y}(\tau_X)$, that is

$$e_{\tau_Y}(\tau_X) = e_{\tau_{Y_2}}\left[e_{\tau_{v_1}}(\tau_X)\right]. \quad (3)$$

Because the true score defined in STST is the proportion-correct true score, the obtained Form X to Form Y conversion function (Equation 3) is for proportion-correct true scores, and the equivalents on the number-correct score metric are found by multiplying the Form Y equivalent proportion-correct true scores by the number of items on Form Y, K_Y. Because true scores of examinees are never known, observed scores are used in place of number-correct true scores as in IRT true score equating. For Form X scores outside the range of possible true scores on Form X, a linear interpolation is implemented.

A variety of STST models can be used for chained beta true score equating under the CINEG design. Because preliminary findings suggest that the assumed true score distribution has a more prominent effect on the equating results, the current study focuses on the two-parameter beta binomial (Beta2) true score method and the four-parameter beta compound binomial (Beta4) true score method.

Equating Methods Used for the Real Data Analysis

Five equating methods were included in the real data analysis, including FE, CE, IRT observed score, IRT true score, Beta2, and Beta4 methods.

Study Factors for the Simulation

Three factors were manipulated for the simulated data analysis, including sample size, group difference, and equating methods. Two sample sizes, 1,000 and 3,000, were considered. Group difference was measured by varying the mean of the population ability distributions from which the two nonequivalent groups were drawn. Five levels of group difference were examined: the population distribution for the old form (population 2) was fixed as $N(0, 1)$, and the population distribution for the new form (population 1) were $N(-0.25,\ 1)$, $N(-0.1,\ 1)$, $N(0, 1)$, $N(0.1, 1)$, and $N(0.25, 1)$. The equating methods considered were FE, CE, IRT true and observed score methods, and Beta2 and Beta4 methods. In total, ten (=2*5) simulation conditions were examined for six equating relationships.

Evaluation Criteria

The criterion equating relationships were defined as IRT observed score equating using the item parameters and population ability distributions assuming the 3PL model is the true model. Because the criterion favors IRT methods, IRT equating methods were not included in the simulated data analysis. It should be noted that using IRT observed score equating as a criterion equating may favor the observed score equating methods. It is also possible to use IRT true score equating as the criterion equating, which would favor the true score equating methods. Therefore, it is virtually impossible in a simulation study to have a criterion that does not advantage or disadvantage one or more methods to some degree. To evaluate the equating results for traditional and beta equating results, conditional absolute bias (CAB), conditional standard error of equating (CSEE), and conditional root mean squared error (CRMSE) were computed using the following equations:

$$CAB(x) = |\bar{e}_Y(x) - e_Y(x)|, \tag{4}$$

where

$$\bar{e}_Y(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{e}_{Yr}(x), \tag{5}$$

$$CSEE(x) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}[\hat{e}_{Yr}(x) - \bar{e}_Y(x)\ ]^2}, \tag{6}$$

$$CRMSE(x) = \sqrt{CAB(x)^2 + CSEE(x)^2}. \tag{7}$$

In Equations 4, 5 and 6, R is the number of replications, which is 500 in this study. In one replication, a specific number of examinees were simulated from the assumed population 1 and 2 distribution, respectively, and then 0/1 responses were generated for each examinee based on the true item parameters. The equating results for the replication were produced using considered equating methods and the generated data. $e_Y(x)$ is the criterion equated score at score x; $\hat{e}_{Yr}(x)$ is the new form equated score for x at replication r; and $\bar{e}_Y(x)$ is the mean equated score for x over $R = 500$ replications. The aggregate evaluation indices including weighted average bais (WAB), weighted standard error of equating (WSEE) and weighted root mean square error (WRMSE), were the weighted averages of the conditional evaluation indices, where the weights were the relative frequencies of scores of the original Form X group.

Results

The real data analysis results are presented in Figure 2. The two traditional methods, FE and CE, yield bumpy equating relationships, and the trends of these two relationships are similar across the score range. At the lower scores (from 0 to 10), where few examinees scored and linear interpolation was used, the trends of the two methods deviated from other equating methods dramatically. The two IRT equating relationships are smooth and generally follow the trend of FE and CE. For the majority of score points (from 20 to 60), the IRT true and observed score equating generate similar results. The observations of the trends for traditional and IRT equating relationships are consistent with previous research. The Beta2 true score method produces smooth equating results across the whole score range. The trend for the Beta2 method approximates the trends of traditional methods at the lower and middle score range and the trends of IRT methods at the higher score range. The differences in magnitudes of the equivalent scores produced by Beta2 and other equating methods are mostly within 1 score point across the score scale. The equating relationship for the Beta4 true score method shows an irregular pattern, which differs from the Beta2 method as well as other equating methods at the lower and higher ends (from 0 to 10 and from 53 to 60). The maximum difference in equivalent scores at the higher end differs from the traditional methods by over 1.5 points. From Figure 2, it is observed that the Beta4 true score method generates a different equating relationship than the other methods.

The simulation results are presented in Figures 3-7 to evaluate the accuracy of Beta2 and Beta4 equating compared to other equating methods. The conditional evaluation statistics

averaged over 500 replications for the condition of sample size of 1,000 and group difference of 0.1 are presented in Figures 3-5. It is evident that the Beta2 method has very little bias across the entire score range and FE and CE show some bias at the lower range, whereas the Beta4 method has moderate bias throughout the entire score scale. The Beta2 method yields the smallest CSEE, followed by FE and CE, which have larger SEE at lower score points. The Beta4 method produces the highest SEE across the score range compared to the other three methods. The same pattern is observed for the CRMSE, which represents the total error. Overall, for the conditional statistics example shown in Figures 3-5, Beta2 produces the most accurate equating results across the score range, while Beta4 produces the least accurate equating results.

Figures 6 and 7 summarize the effect of group difference on equating accuracy for sample size of 1,000 and 3,000, respectively. The patterns for the examined equating methods are consistent for the two sample-size conditions, with the magnitude of WSEE and WRMSE decreasing as sample size increases. As group ability difference increases, WAB of FE and CE increases, while WAB of Beta2 doesn't show much difference, suggesting that Beta2 is not much affected by group differences. Although Beta4 also appears to be unaffected by group differences, it is the least accurate of the four methods, especially in terms of WSEE and WRMSE.

Discussion

It is important for testing organizations to choose equating methods that provide accurate results. In contrast to the large amount of literature on traditional and IRT equating, very few studies address the application of beta equating. Chained beta true score equating has two potential benefits: it is supported by a solid psychometric model, and there are considerably fewer parameters to be estimated compared to IRT models. To gain a better understanding of chained beta true score equating, this study offered intensive investigation of chained beta true score equating for the CINEG design. Real and simulated data were used to evaluate the accuracy of Beta2 and Beta4 methods under different sample size and group ability distribution conditions. The results show that the Beta2 method generates a smooth equating relationship, producing more accurate equating results compared to FE and CE. Additionally, the Beta2 method was not affected by group differences. The Beta4 method generated very different equating relationships than the other methods and yielded the least accurate equating results. One

possible reason of the poor performance of Beta4 method could be the estimates of lower and upper ends of the four-parameter beta distribution.

There are a few limitations of this study. First, the findings were obtained by defining the IRT observed score method as the criterion equating, which is a commonly used method in equating literature for investigating the accuracy of equating methods. However, the conclusions may be different if other criteria were used, so caution should be made in generalizing the results. Another limitation is that only one set of real data was used in this study, so future studies should examine the chained beta true score equating methods using other tests of different length and different proportion of common items to examine if the chained beta true score methods generate consistent equating results for different forms of real data.

This study mainly focused on the CINEG design, so future studies should be conducted on the application of beta equating methods for the random groups design. In addition, chained beta observed score equating can be examined and compared with existing methods. It is also worthwhile to compare chained beta methods with smoothed equipercentile equating methods and the modified FE method to evaluate equating accuracy.

References

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

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. (Available from the web site: http://www.uiowa.edu/~casma).

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

Hanson, B. A. (1991). *Method of moments for the four-parameter beta compound binomial model and the calculation of classification consistency indexes*. (ACT Research Report 91-5). Iowa City, IA: American College Testing.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report No. 94-4). Iowa City, IA: American College Testing.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, *27*, 59–72.

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

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

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 30,* 239-270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing. (An empirical Bayes estimation problem.). *Psychometrika*, *34*, 259–299.

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

Wang, T., Lee, W., Brennan, R. L., & Jing, S. (in press). *Assessing equity property for equating under strong true score and IRT frameworks* (CASMA Research Report No. 53). Iowa City, IA: University of Iowa.

Table 1

Descriptive Statistics and Reliability for Form X and Form Y of Test 1

Test	Form	Length	Mean	SD
Test 1	X	60	34.09	11.39
	V	12	6.96	2.79
	Y	60	31.47	11.92
	V	12	6.64	2.76

Table 2

Summary Statistics of the Item Parameters for Form X and Form Y of Test 2

Test	Form	Length	***a*-parameter**		***b*-parameter**		***c*-parameter**	
			Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Test 2	X	50	0.76	0.27	-0.32	0.94	0.15	0.22
	V	12	0.71	0.25	-0.33	0.99	0.12	0.18
	Y	50	0.78	0.30	-0.35	0.99	0.12	0.20
	V	12	0.73	0.28	-0.36	0.99	0.11	0.17

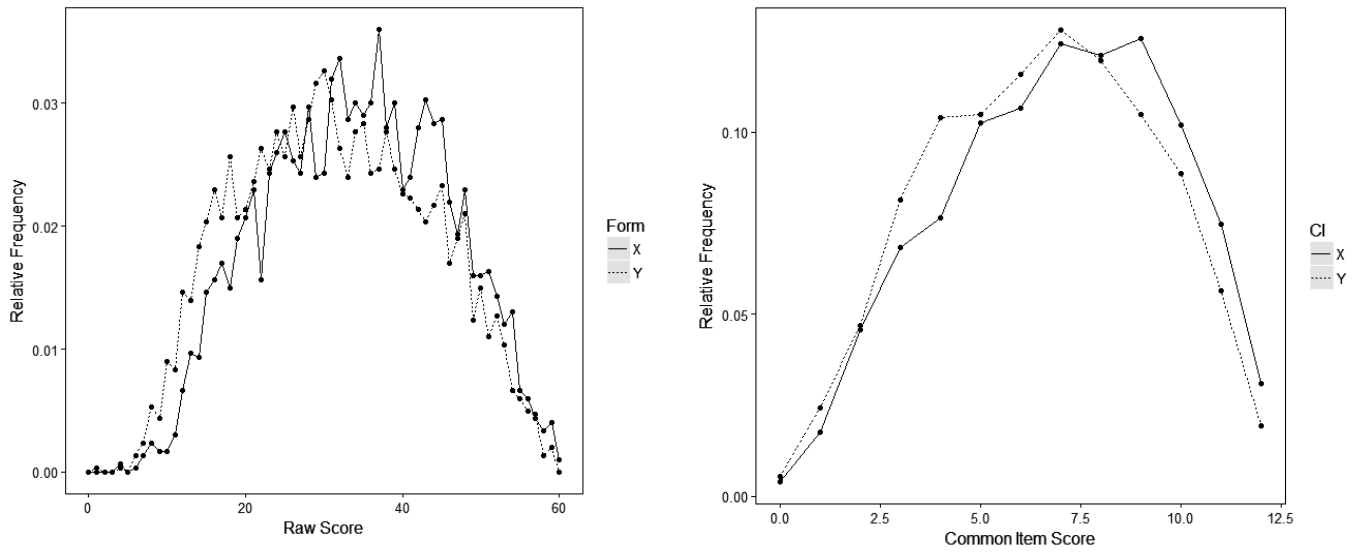


Figure 1. Observed score distributions for Test 1.

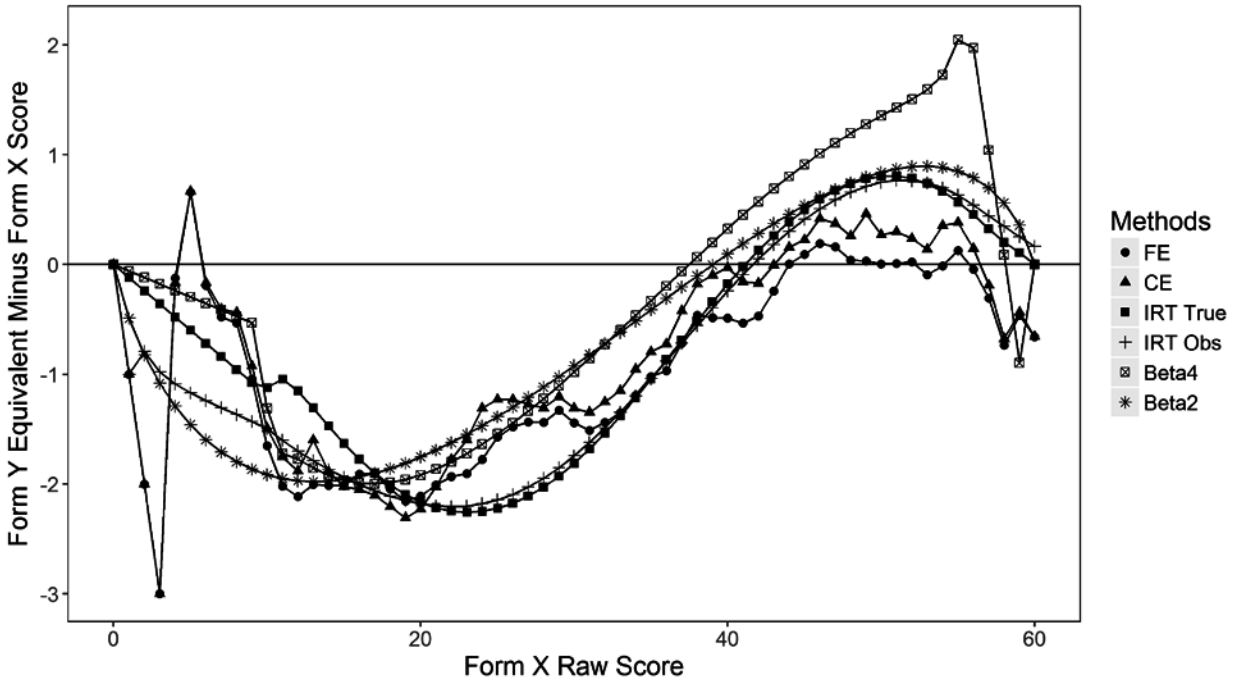


Figure 2. Real data analysis results: equating relationships expressed as differences for Form Y equivalent minus Form X score.

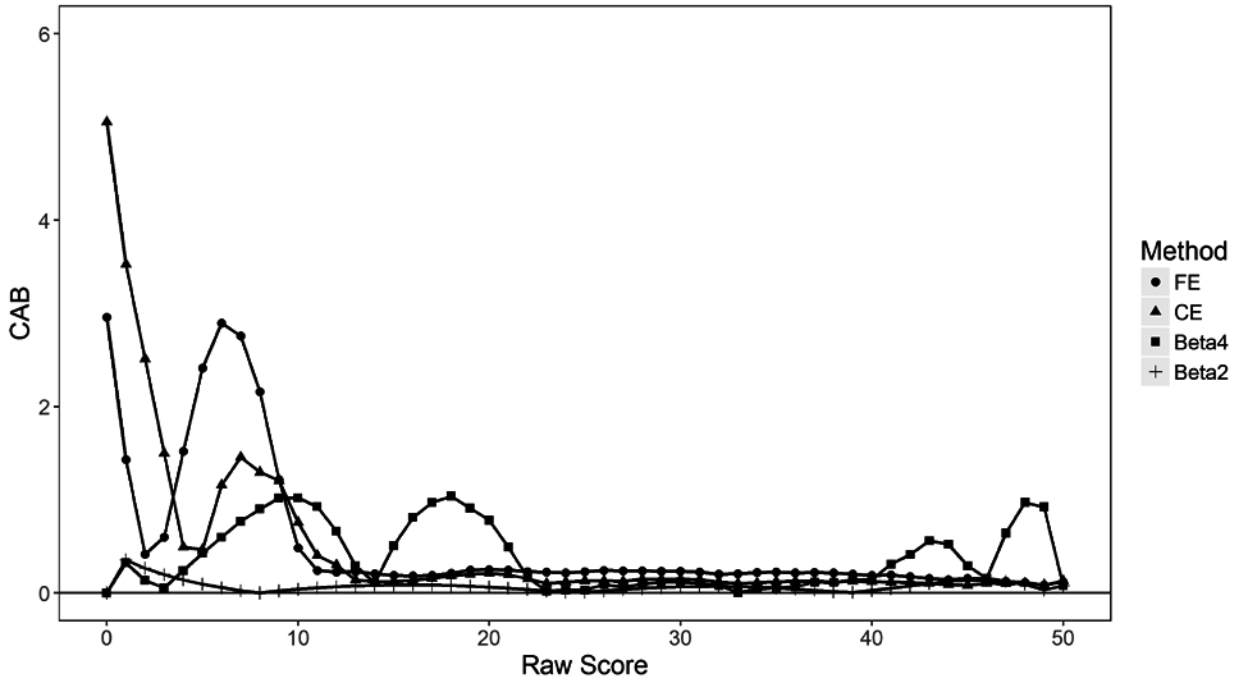


Figure 3. Simulation results: CAB averaged over 500 replications for sample size of 1,000 and group difference of 0.1.

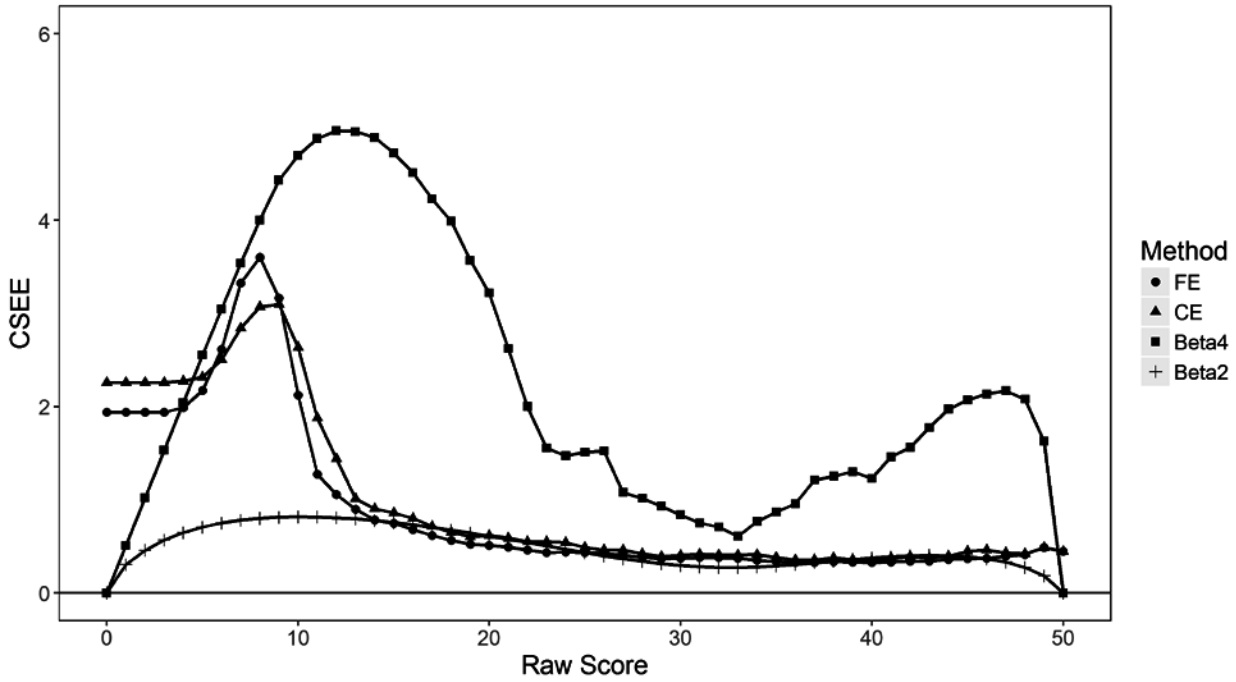


Figure 4. Simulation results: CSEE averaged over 500 replications for sample size of 1,000 and group difference of 0.1.

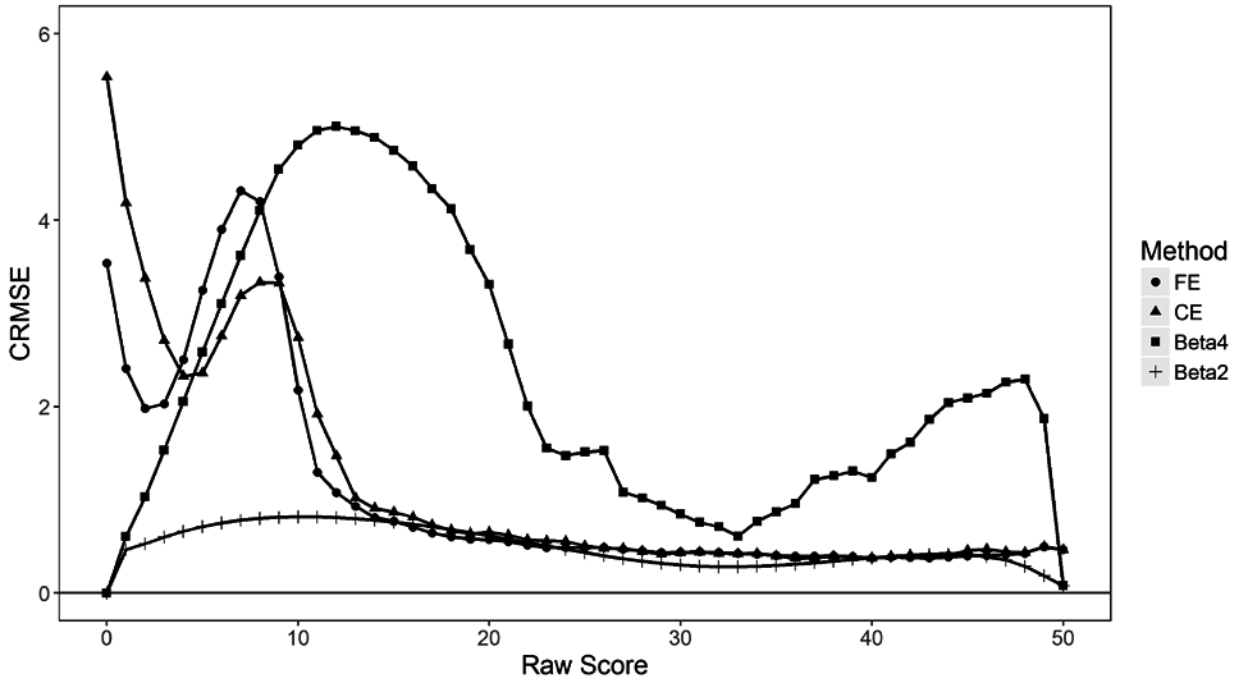


Figure 5. Simulation results: CRMSE averaged over 500 replications for sample size of 1,000 and group difference of 0.1.

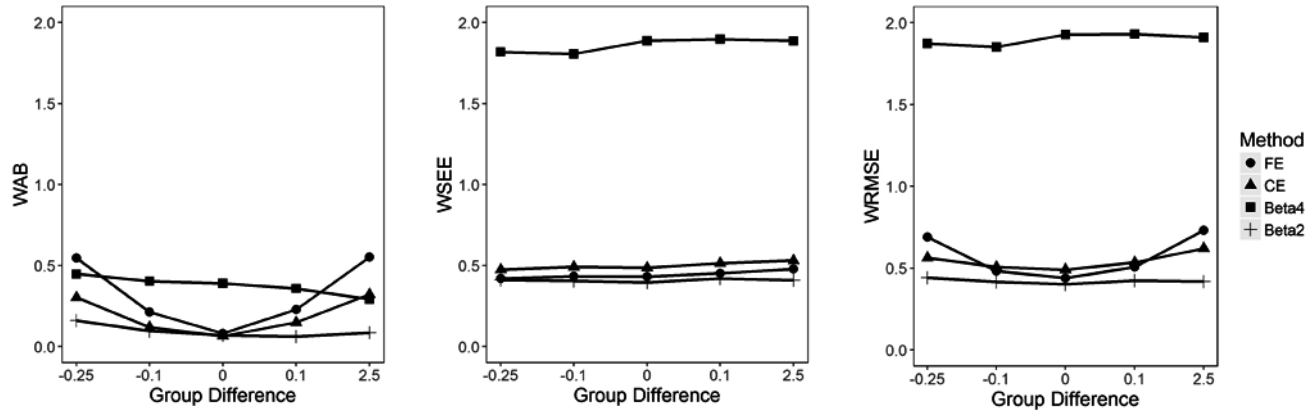


Figure 6. Simulation results: WAB, WSEE and WRMSE for sample size of 1,000.

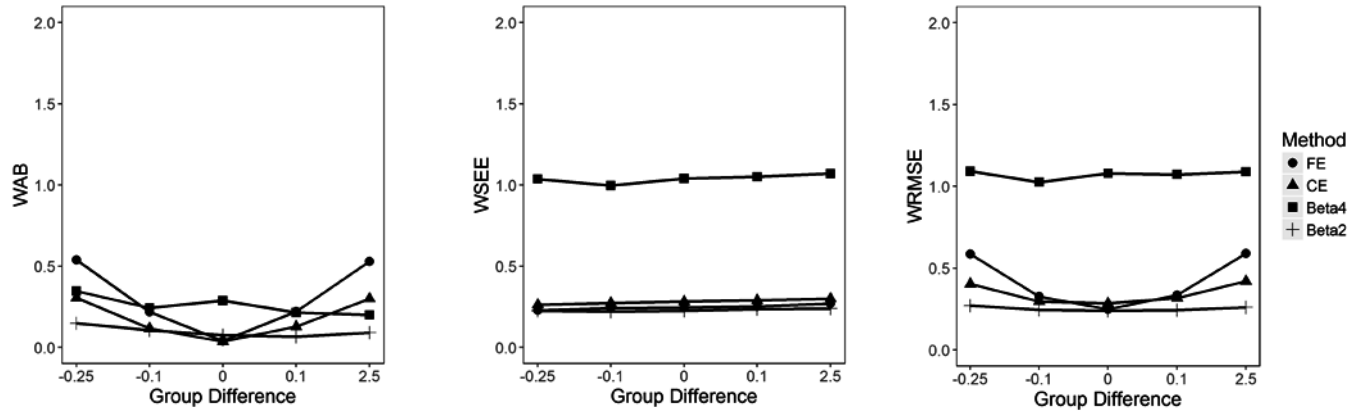


Figure 7. Simulation results: WAB, WSEE and WRMSE for sample size of 3,000.

Chapter 7: Exploring Score Dimensionality for Mixed-Format Tests Using Factor Analysis and Item Response Theory

Mengyao Zhang

National Conference of Bar Examiners, Madison, WI

Michael J. Kolen and Won-Chan Lee

The University of Iowa, Iowa City, IA

Abstract

Dimensionality assessment provides test developers and users with empirical knowledge of how the intended test structure is reflected by actual test data, which is critical in validating interpretations and uses of test scores. The use of mixed-format tests results in a more complex dimensional structure compared to single-format tests and poses methodological challenges to dimensionality assessment. The purpose of this study was to review and compare two promising methods that are suitable for exploring the dimensional structure for mixed-format tests. The specific methods considered included: item-level exploratory factor analysis (item-level EFA) and multidimensional IRT cluster analysis (MIRT cluster analysis), which rely on factor analysis and IRT definitions of dimensionality, respectively, and employ different procedures to determine the number of dimensions, cluster items, and interpret results. Mixed-format test data from three Advanced Placement exams were used. Results revealed inconsistencies in the number of dimensions and the clustering solution between the two methods, and showed how dimensionality assessment results were associated with the subject area, form, and sample size factors. In addition, item-level EFA and MIRT cluster analysis were used to investigate potential sources of multidimensionality for the selected exams. No clear evidence concerning the format effect was found, whereas content distribution and testlet dependence appeared to be related to the dimensional structure.

Exploring Score Dimensionality for Mixed-Format Tests Using Factor Analysis and Item Response Theory

An effective educational test reflects important differences among examinees on one or more dimensions it intends to measure, such as reading comprehension, mathematical reasoning, and chemistry laboratory skills (Yen & Fitzpatrick, 2006). Dimensionality assessment provides test developers and users with empirical knowledge of how these intended dimensions are reflected by the actual test data and whether or not some unexpected dimensions emerge, for example, due to inappropriate reading load or speededness. The dimensional structure of the data is one of the critical pieces of evidence used to validate interpretations and uses of test scores.

Dimensionality could be evaluated in both exploratory and confirmatory manners (Reckase, 2009; Svetina & Levy, 2014). Exploratory dimensionality assessment, which is the focus of this study, is typically employed when there is no clear hypothesis or evidence concerning the dimensional structure of the given data. It has been routinely conducted, either alone or combined with confirmatory dimensionality assessment, in operational testing programs to check the alignment of the actual dimensionality with the intended dimensionality (e.g., Jang & Roussos, 2007; Zwick, 1987). Exploratory dimensionality assessment also often serves as an integrated part of a preliminary analysis before studying other psychometric procedures such as equating and linking (e.g., Brossman & Lee, 2013).

In recent years, many state and national educational programs have adopted mixed-format tests. A typical mixed-format test contains both the multiple-choice (MC) and free-response (FR) items in order to combine strengths of different item formats and mitigate possible weaknesses of using either MC or FR items alone. The widespread use of mixed-format tests, however, complicates dimensionality assessment in several ways.

Conceptually, data from mixed-format tests tend to have a more complex dimensional structure compared to those from single-format tests. A unique source of multidimensionality associated with mixed-format tests is the item format, and previous research studies on whether varying item format introduces extra dimensions, known as the format effect, have shown mixed findings (e.g., Bennett, Rock, & Wang, 1991; Bridgeman & Rock, 1993; Hohensin & Kubinger, 2011; Thissen, Wainer, & Wang, 1994; Traub & Fisher, 1977). Data from mixed-format tests are also vulnerable to the same potential sources of multidimensionality that affect single-format

tests, such as the content distribution, passage or stimulus dependence in testlets, speededness, rater effects, inappropriate external interference, and so on (Yen, 1993).

A methodological challenge facing dimensionality assessment of mixed-format test data stems from different types of item scores used with MC and FR items. Typically, MC items are dichotomously scored, whereas FR items are polytomously scored. Thus, an ideal method for analyzing mixed-format test data should be able to handle dichotomous and polytomous item data simultaneously, which eliminates some widely used procedures developed for MC-only tests (Svetina & Levy, 2012). Even for those methods 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 data from operational mixed-format tests.

The potential problems concerning examination of score dimensionality for mixed-format tests motivated the current study. The primary purpose of this study is to apply two promising dimensionality assessment methods to analyzing the dimensional structures of three Advanced Placement (AP) exams. Each exam is a mixed-format exam that contains both MC and FR items. The two methods considered in this study are: item-level exploratory factor analysis (item-level EFA; Muthén & Muthén, 1998-2012) and multidimensional item response theory cluster analysis (MIRT cluster analysis; Miller & Hirsch, 1992). These methods represent the EFA and item response theory (IRT) perspectives, respectively, which are perhaps the most popular perspectives used to define and assess dimensionality nowadays. Although both methods are capable of exploring the dimensional structures for mixed-format tests, they are different in terms of their definitions of dimensionality, procedures for deciding the number of dimensions and forming item clusters, and interpretations of results. In this study, the performance of these methods was compared under some realistic conditions where the test subject area, form, and sample size varied. These methods were further used to inform sources of multidimensionality. Whether multidimensionality is possibly due to the format effect is of interest, although some other factors affecting dimensionality were also examined in some situations.

Background Information

This section reviews definitions of dimensionality from EFA and IRT perspectives and describes the dimensionality assessment methods used in this study.

Defining Dimensionality

Examination of dimensionality is more challenging than it might seem. A fundamental problem facing researchers and practitioners is: what is meant by dimensionality? Hattie (1981) first clarified the term unidimensionality and distinguished it from a set of related terms. Later, various operational definitions of dimensionality have been developed, serving as the basis for different dimensionality assessment methods that are currently in use in practice. In this study, two operational definitions of dimensionality were considered: one is from an EFA perspective, and the other one is from an IRT perspective. The EFA and IRT definitions of dimensionality are perhaps the most commonly used operational definitions of dimensionality nowadays.

Dimensionality has long been analyzed using factor analysis (e.g., Hattie, 1985; Reckase, 2009; Stone & Yeh, 2006; Velicer, Eaton, & Fava, 2000). Principal factor analysis is one of the EFA procedures that have been typically applied to assessing dimensionality when strong prior beliefs about a test's internal structure are lacking. It 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 generally employed to enhance the interpretations of results. From an EFA perspective, the number of factors retained equals the number of dimensions. The data are considered to be unidimensional when only one factor is kept, and otherwise some degree of multidimensionality emerges. Dimensional structure underlying the data corresponds to a particular factor solution. 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.

Traditional EFA treats observed variables as continuous, whereas item scores from mixed-format tests are a mixture of dichotomous and polytomous variables. It has been shown that the use of Pearson's product-moment correlation — assuming that observed categorical variables are continuous — tends to distort the actual relationship between item scores and leads to the presence of spurious dimensions and biased estimation of factor loadings (e.g., Hattie, 1985; Olsson, 1976). A possible solution is to replace the Pearson correlation with the polychoric correlation that is defined as the correlation between two continuous latent variables underlying

item scores. Improvements in computational abilities in recent years facilitate the application of polychoric correlations in dimensionality assessment. However, especially with small sample sizes, a polychoric correlation matrix often fails to be positive definite (PD). Eigenvalues of such a matrix are not always positive, making EFA that is originally built on Pearson's correlations not strictly applicable. Either specialized model estimation methods or smoothing methods (i.e., used to transform a non-PD matrix to a PD matrix) should be used.

Alternatively, dimensionality could be defined using IRT, and this perspective is attractive to many researchers and practitioners because it offers clear specifications of the relationship between item scores and the so-called latent traits (e.g., Nandakumar, 1991; Stout, 1987, 1990). A traditional IRT definition of dimensionality is the minimum number of latent traits required for a locally independent and monotone model (Stout, 1990). Specifically, local independence implies that, after controlling for underlying latent traits, item responses are either mutually independent or pairwise uncorrelated, depending on whether strong or weak local independence is considered. Monotonicity indicates that the probability of correctly answering an item changes monotonically with values of latent traits. When a single latent trait is sufficient to produce such a model, the data are considered to be unidimensional. If the data are not unidimensional, the number of latent traits required defines the number of dimensions.

For operational tests especially mixed-format tests, perfect unidimensionality where all items strictly measure a single dimension likely rarely occurs. Therefore, only the essential dimensionality of data was considered in this study, which focuses on a number of major or dominant dimensions. The concept of essential dimensionality was formulated mathematically using IRT, which is defined as the minimum number of latent traits required for an essentially independent and weakly monotone model (Junker, 1991; Stout, 1987, 1990). Similar ideas have also been implicitly expressed from the EFA perspective; for example, factors accounting for the majority of observed correlation are kept whereas the other minor factors are ignored (Stout, 1990). In the remainder of this chapter, the two terms *dimensionality* and *essential dimensionality* are used interchangeably without explicit distinction being made.

Assessing Dimensionality

Item-level EFA. Item-level EFA can be used to explore the factor structure of data from mixed-format tests after deciding on a range of the number of dimensions. Mplus (Muthén & Muthén, 1998-2012) was considered in this study because it has been acknowledged as one of

the most popular programs for assessing dimensionality and one of the most flexible packages for handling different types of data (Svetina & Levy, 2012). In Mplus, both the ML and LS estimation methods for conducing EFA are provided, and several different types of rotation methods are available. Although there exists no single "correct" rotation (Sass & Schmitt, 2010), a specific oblique rotation called promax rotation was used in this study, because (a) underlying dimensions, if any, would likely be correlated with each other for a practical mixed-format tests, and (b) promax rotation has been shown to be conceptually simple and computationally efficient (Abdi, 2003).

The output file of Mplus contains estimates of factor loadings, R-squares, and residuals. Some model-fit statistics are also provided, such as model chi-square, root mean square error of approximation (RMSEA), and root mean square residual (RMSR). The model chi-square and RMSEA statistics are generally available under ML and some types of LS estimation, and a smaller value indicates better fit of an EFA model. But model chi-square is sensitive to large sample sizes, making it a less preferable choice for large-scale test data (Kline, 2010). The RMSR statistic that is provided under both ML and LS estimation measures the overall difference between the observed correlation and the correlation predicted by an EFA model, so that a smaller value represents better model fit.

Models with an increasing number of factors are examined sequentially. The following guidelines for choosing between models have been suggested in previous studies (e.g., Kline, 2010; Stone & Yeh, 2006). First, under promax rotation, factor loadings no longer represent correlations between item scores and underlying dimensions, so they should be evaluated along with the corresponding structure coefficient. If both exceed a certain value (e.g., 0.30), the relationship between the item and dimension is considered to be substantial. A dimension is considered to be nontrivial if it is substantially related to more than five items. In this study, the five-substantial-item requirement might sometimes be loosened, because for some tests the number of FR items is less than five. Second, the R-square for an item represents the proportion of observed variances explained by a specific factor solution and is 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 internal structure. Thus, the values of R-squares for individual items might be far from satisfactory under general guidelines. Instead, increments in R-squares by using $(m + 1)$ factors rather than m factors might be more helpful for weighing gain-and-loss when choosing

between models. Third, residuals are expected to be small in absolute value, 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 model chi-square, the RMSEA statistic is more favorable and a rule of thumb is that an RMSEA value less than 0.05 indicates good fit of an EFA model. Last, significance tests are probably misleading given large sample sizes. Solutions with fewer dimensions and easier interpretability are preferable.

MIRT cluster analysis. An IRT approach developed by Miller and Hirsch (1992) can be used to explore and confirm dimensionality of data, which is a combination of MIRT and cluster analysis. The procedure consists of three steps as follows.

The first step is to estimate item parameters of a MIRT model with a specified number of dimensions. In order to capture any possible multidimensionality in the data, the authors suggested using a sufficiently large number according to the researcher's prior experience. There are a number of MIRT models available for dichotomous-scored MC items and polytomously-scored FR items. For this study, the multidimensional three-parameter logistic model (M3PL; Reckase, 2009) was considered for MC items, which is expressed as

$$p_j = \Pr\left(X_j = 1 \middle| \boldsymbol{\theta}\right) = c_j + \left(1 - c_j\right)\frac{\exp\left(\mathbf{a}_j\boldsymbol{\theta} + d_j\right)}{1 + \exp\left(\mathbf{a}_j\boldsymbol{\theta} + d_j\right)}, \quad (1)$$

where multiple latent traits contained in $\boldsymbol{\theta}$ together determine how likely an examinee answers item j correctly; $\mathbf{a}_j$ is a transposed vector of slope parameters; d_j is the intercept; and c_j is the pseudo-guessing parameter. Equation (1) also implies the compensatory nature of the M3PL model: a low value on one latent trait could be compensated by high value(s) on one or more of the other latent traits. Compensatory models have been applied extensively to dimensionality assessment, as they seem to better reflect the actual cognitive processes observed in many educational tests and do not have serious theoretical or computational problems (e.g., Mroch & Bolt, 2006; Nandakumar, 1994).

The multidimensional graded response model (MGR; Muraki & Carlson, 1995) was considered for FR items, which begins with the expressions of cumulative category response functions,

$$p_{j0}^{*} = \Pr\left(X_j \geq 0 \middle| \boldsymbol{\theta}\right) = 1; \quad (2)$$

$$p_{jk}^{*} = \Pr\left(X_j \geq k \middle| \boldsymbol{\theta}\right) = \frac{\exp\left(\mathbf{a}_j\boldsymbol{\theta} + d_{jk}\right)}{1 + \exp\left(\mathbf{a}_j\boldsymbol{\theta} + d_{jk}\right)}, k = 1, \ldots, K_j - 1.$$

The parameters in Equation (2) are similar to those in Equation (1), except that the intercept d_{jk} is assigned to response categories from 1 to $\left(K_j - 1\right)$ for a polytomously-scored FR item. The MGR model is then defined as the difference between cumulative category response functions:

$$\begin{aligned} p_{jk} &= \Pr\left(X_j = k \middle| \boldsymbol{\theta}\right) = p_{jk}^{*} - p_{j(k+1)}^{*}, k = 0, \ldots, K_j - 2; \\ p_{j(K_j-1)} &= \Pr\left(X_j = K_j - 1 \middle| \boldsymbol{\theta}\right) = p_{j(K_j-1)}^{*}. \end{aligned} \tag{3}$$

The MGR model also allows for high ability on one latent trait to compensate for low ability on another latent trait.

In the second step, the angular distance between every pair of items j and j' is calculated based on estimated item parameters in the first step, which is defined as

$$a_{jj'} = \arccos \frac{\mathbf{a}_j\mathbf{a}_{j'}^{\mathrm{T}}}{\sqrt{\mathbf{a}_j\mathbf{a}_j^{\mathrm{T}}} \times \sqrt{\mathbf{a}_{j'}\mathbf{a}_{j'}^{\mathrm{T}}}}, \tag{4}$$

where $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ are vectors of slopes for items j and j', respectively. Angular distances for all possible item pairs constitute the angular distance matrix.

In the last step, a hierarchical cluster analysis, such as the complete linkage cluster method used in Miller and Hirsch (1992), is conducted, where the angular distance matrix is treated as a dissimilarity matrix.

Determining the number of clusters might be the most challenging part of this method, because objective criteria combined with subjective judgments are required to find a stopping point. Milligan and his collaborators conducted extensive comparative studies on the evaluation of criteria for cluster analysis. Five of 30 criteria have been suggested in Milligan and Cooper (1985) for showing superior capabilities of recovering the true cluster structure, including the Caliński and Harabasz index (CH; Caliński & Harabasz, 1974), Duba and Hard index (DH; Duba & Hart, 1973), C-index (Hubert & Levin, 1976), Gamma index (Baker & Hubert, 1975), and Beale index (Beale, 1969). In particular, the CH and C-index examine squared distances between and within clusters, and the Gamma index checks comparisons between within-cluster dissimilarities and between-cluster dissimilarities. For these indices, the largest value indicates

the optimal number of clusters. The DH and Beale indices assess whether a cluster should be divided into two smaller clusters, and the optimal number of clusters is found once the hypothesis of one cluster is rejected. In theory, the CH, C-, and Gamma indices cannot be used to assess if there is only one cluster (i.e., unidimensionality), whereas the DH and Beale indices can. As shown in Milligan and Cooper (1985), the ordering of criteria for cluster analysis might vary when different data were used, but the best indices seemed unlikely to show extremely poor performance. Among the five best criteria as indicated above, the CH index consistently provided the best cluster recovery for simulated data containing two, three, four, or five clusters. The remaining four criteria performed worse when the data contained two clusters. When errors occurred the DH and Beale indices suggested too few clusters, whereas the C- and Gamma indices suggested too many clusters. It should be noted that in the literature these criteria were examined assuming the Euclidean distance, not the angular distance used in MIRT cluster analysis. The performance of these criteria in this new context is still open to question.

Miller and Hirsch (1992) also recommended that users pay attention to the overall dimensional pattern of the data and attach content or other meaningful interpretations to clusters, rather than simply stopping at a statistically optimal point. In addition, Mroch and Bolt (2006) emphasized usefulness of dendrograms from cluster solutions in the discovery of underlying multidimensional structure.

Method

Data

Data used in this study were from three AP exams administered in 2011 and 2012, including English, Spanish, and Chemistry. Three factors were primarily considered in the selection of exams: subject, form, and sample size. Specifically, the selected exams represent the subject areas of language and science. Within the language area, both English and Spanish were included in this study because they focus on mastery of a native language and a foreign language, respectively. Assessments of native and foreign languages tend to have different purposes, impacting the internal structure of scores collected from those exams. Compared to English, Spanish also measures more types of language skills and contains more complex integrated tasks, which could further complicate dimensionality assessment. For each exam, the 2011 and 2012 forms were used to examine if dimensionality assessment methods perform similarly across groups of examinees taking parallel forms. The selected exams did not undergo

any major redesign during this period. Thus, those forms can be viewed as being comparable, and dimensional structure of data from those forms was expected to be similar. Last, each exam was administered to over 15,000 examinees. Therefore, samples of different sizes could be generated, allowing for investigation of how the sample size affects dimensionality assessment.

Additional factors were taken into account when selecting exams. For example, the test length varied across exams: English had only 54 or 55 MC items and 3 FR items, whereas Chemistry consisted of 75 MC items and 6 FR items. For the two language exams, all the MC items were grouped into testlets sharing the same stimuli; for Chemistry, only a few MC items were in testlets. Different types of FR items were also used to better fulfill the proposed test purposes for selected exams, for instance, short answer, synthesis essay, and speaking prompts under interpersonal and presentational scenarios.

For this study, MC items were number-correct scored as 0 or 1, and FR items were polytomously scored on integer scales. 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, section weighting was not considered as an investigation factor in this study, and only item scores and total scores were used. Total scores were calculated as sums of MC and FR score points without any weights. In addition, some FR items with more than 10 response categories were rescored before dimensionality assessment, in order to make dimensional solutions from using different methods more comparable. After rescoring, all FR items on each form of each exam had no more than 10 response categories, while the resulting data only slightly deviated from the original data with regard to their statistical characteristics. Descriptive statistics for the data used in this study are summarized in Table 1.

The original exam data were collected under the common-item nonequivalent groups (CINEG) design: the 2011 form served as the old form and the 2012 form was the new form constructed to be comparable to the old form in terms of content and statistical characteristics and sharing a certain number of MC items with the old form. Given this relationship between two forms of each exam, two random samples of N = 3,000 were generated from populations who took the 2011 and 2012 forms, respectively, allowing for investigation of form difference in dimensionality assessment. The value of 3,000 is a relatively large sample size for performing the selected methods considered in this study. From the population who took the 2011 form, two additional random samples of N = 500 and 1,000 were generated so as to inspect the effect of

sample size on dimensionality assessment. The value of 500 is considered to be a small sample size. This size might approximately satisfy the requirements of EFA, but the estimated polychoric correlations might contain substantially large sampling error and the matrix is typically non-PD. For MIRT cluster analysis, a small sample size might also cause incorrect estimation of IRT model parameters. The value of 1,000 is considered to be slightly above the minimum required sample size.

Exploring the Dimensional Structure

Item-level EFA. When running item-level EFA, the range of the number of factors to be considered needs to be input by the researcher. In this study, the minimum number of factors was always set to one, implying that unidimensionality holds. The maximum number of factors was decided based on results of an earlier empirical study on these AP exams (Zhang, Kolen, & Lee, 2014). Specifically, the maximum numbers of factors were assigned to four for datasets from English and Chemistry and six for datasets from Spanish.

Once the range of the number of factors was decided, item-level EFA was conducted on polychoric correlations using Mplus. Item scores were read directly into Mplus for analysis. The default WLSMV estimator for categorical variables was chosen, which is a robust weighted LS estimation method. Promax rotation was used to improve interpretations of results. After running Mplus, factor model selection was made according to the overall model-fit statistics, factor loadings and factor structure coefficients, correlation residuals, R-squares and R-square increments, as well as the principle of parsimony and the interpretability of the factor solution. Specific model-fit statistics checked in this study included model chi-squares for the single model and model comparison, RMSEA, and RMSR.

MIRT cluster analysis. Implementation of MIRT cluster analysis required the combination of two software packages. Item parameters of MIRT models were estimated by flexMIRT (Cai, 2013) and passed to R for cluster analysis. The numbers of dimensions assumed in MIRT calibration were decided similarly to those in item-level EFA: MIRT models with four dimensions were used to fit datasets from English and Chemistry, and MIRT models with six dimensions were used to fit datasets from Spanish. The M3PL model was chosen for MC items, and the MGR model was chosen for FR items. The MIRT models were estimated using the default EM algorithm in flexMIRT. The calibration settings can be seen in Table 2, which have also been used in previous research studies concerning calibration of AP exams (Kolen & Lee,

2014). In order to solve several indeterminacies in MIRT calibration, $\boldsymbol{\theta}$ was assumed to follow the default multivariate normal distribution in flexMIRT, and any slope having the dimension number greater than the item number was fixed to 0. For example, slopes associated with the first item on the second and higher dimensions were forced to be 0, slopes associated with the second item on the third and higher dimensions were forced to be 0, and so forth. Such constraints have been typically used in other calibration programs to calibrate MC items, such as TESTFACT and NOHARM (Reckase, 2009).

Based on estimated item parameters, especially estimated slopes, the angular distance matrix was calculated in R, serving as the dissimilarity matrix in cluster analysis. The five criteria described in the previous section — the CH, DH, C-, Gamma, and Beale indices — were used to determine the optional number of clusters. The NbClust package in R (Charrad, Ghazzali, Boiteau, & Niknafs, 2014) was used to calculate relevant statistics. Possibly meaningful interpretations were also taken into account when opting for a particular dimensional solution. Extensive graphical representations were checked, such as dendrograms at varying levels.

Evaluating the Format Effect

Item-level EFA and MIRT cluster analysis were also used in this study to investigate the format effect. Results related to the largest datasets (i.e., N = 3,000) from the 2011 form of the selected exams were used. Item-level EFA produced factor solution comparable to those clusters of items estimated by MIRT cluster analysis.

Evaluation Criteria

A primary purpose of this study was to compare the performance of item-level EFA and MIRT cluster analysis in revealing the dimensional structure underlying the data, and how the subject area, form, and sample size factors are associated with the level of consistency between these methods. The level of consistency between dimensional structure estimated by the two methods was evaluated as follows for each condition (similar to the matching coefficient in Mroch & Bolt, 2010).

First, a two-way table was constructed to store proportions of consistent and inconsistent structural identifications between the two methods, as illustrated in Figure 1. Specifically, for any J-item exam, there are $J(J-1)/2$ item pairs (the order of items in a pair does not matter). Any item pair must contribute to one of the four proportions: (a) proportions of item pairs

assigned to different dimensions (i.e., items apart) by the two methods, denoted as φ_{00}; (b) proportion of item pairs assigned to the same dimension (i.e., items together) by the two methods, denoted as φ_{11}; (c) proportion of item pairs assigned to different dimensions by the first method but assigned to the same dimension by the second method, denoted as φ_{01}; and (d) proportion of item pairs assigned to the same dimension by the first method but assigned to different dimensions by the second method, denoted as φ_{10}.

Next, a consistency index was calculated as

$$\varphi = \varphi_{00} + \varphi_{11}. \quad (5)$$

Possible values of φ range from 0 to 1. As the value of φ approaches 1, dimensional structure estimated by the two methods becomes more and more alike.

For example, suppose an exam contains 10 items, denoted as Item 1 through Item 10. By using item-level EFA, Items 1–5 are clustered around one dimension, and Items 6–10 are clustered around another dimension. However, MIRT cluster analysis detects three dimensions measured by Items 1–3, Items 4–7, and Items 8–10, respectively. Recall that every item pair must contribute to one of the four proportions defined above. Consider the pair of Item 1 and Item 2. These items are assigned to the same dimension according to either method, so this pair contributes to φ_{11}. Now consider the pair of Item 5 and Item 6. These items are assigned to different dimensions by item-level EFA but assigned to the same dimension by MIRT cluster analysis. Thus, this pair contributes to φ_{01}. After examining all possible 10×9/2 = 45 pairs, a two-way table is completed, as shown in Figure 2, and φ is approximately 0.65.

Results

Descriptive Statistics

Table 3 shows the score range and moments of total scores for each dataset used in this study. After taking into account the variability of total scores, only small differences were found among the mean score of the three samples from the 2011 population, which is largely attributed to random sampling error, and the difference between the mean score of the 2011 and 2012 samples of N = 3,000 was not substantial. The distribution of total scores across different samples for each exam was also similar. Common-item effect sizes were further calculated for the 2012 form minus the 2011 form common-item scores, indicating how groups of examinees taking different forms differ from each other. A positive value indicates that examinees taking the 2012 form are more able than those taking the 2011 form. For English, Spanish, and

Chemistry, values of those statistics were 0.009, 0.058, and 0.023, respectively, which are not shown in the table. The absolute differences between groups of examinees were all below 0.10 and might be considered to be relatively small. All of these statistics suggested that the data used in this study are reasonable for assessing whether dimensional results were stable across parallel forms of the same exam and across random samples from the same population.

For each dataset, reliability coefficients using Cronbach's alpha (Cronbach, 1951) were estimated separately for MC and FR sections, and reliability of total scores was evaluated using the stratified alpha coefficient by viewing the exam as being composed of the MC and FR "strata." Values of these reliability coefficients are presented in Table 4. As shown in the table, the MC section typically had higher reliability than the FR section. Reliability of total scores for all the datasets were around or above 0.90, implying that levels of internal consistency of total scores were relatively high.

A rough examination of format effects was conducted by checking disattenuated correlations between MC and FR items for the selected exams. These disattenuated correlations are also displayed in Table 4. Disattenuated correlations between MC and FR scores for English were almost always the lowest (around 0.80), followed by those for Spanish (around 0.82), suggesting that MC and FR items on two language exams might measure somewhat different dimensions. By contrast, disattenuated correlations between MC and FR scores for Chemistry were above 0.95. For this exam, separate dimensions associated with the format might not be obvious.

Item characteristics were checked because irregular items can hinder performance of some dimensionality assessment procedures. For example, MIRT models might not converge when there are many items with extreme difficulty or discrimination. In this study, the average item score divided by the maximum possible item score was used to represent item difficulty. Values of this statistic range from 0 to 1: the higher the value, the easier the item. For MC items, this value is equivalent to the p-value. Biserial or polyserial correlation between item score and rest score (i.e., total score minus the given item score) was used to indicate item discrimination. Values of this statistic range from -1 to 1, though negative values are rarely seen in well-designed educational tests. A higher value indicates greater power of an item to distinguish between lower- and higher-performance examinees. Means and standard deviations of the above item difficulty and discrimination statistics for MC and FR items are presented in Tables 5 and 6.

Items were flagged as irregular if any of the following occurred: (a) item difficulty was lower than 0.10; (b) item difficulty was higher than 0.95; or (c) item discrimination was lower than 0.10. As shown in these tables, despite there being few irregular MC items, items on the selected exams had adequate levels of difficulty and reasonably good discrimination. A major limitation of these item statistics is that they are sample dependent, meaning that values of these statistics change as different samples of examinees are used. However, these statistics usually rely on weak assumptions, especially compared to IRT item parameters. Item parameters of a traditional unidimensional IRT model were intentionally avoid, because the unidimensionality assumption, one of the most fundamental assumption underlying such a model, could not be made prior to dimensionality assessment.

Dimensional Structure

In this study, the dimensional structure of each dataset was estimated by item-level EFA and MIRT cluster analysis. Items on a form were grouped into several clusters according to the dimensions influencing their scores. The remainder of this section is further divided into three parts. The first two parts describe results from using item-level EFA and MIRT cluster analysis individually. The last part discusses to what extent the two methods provided consistent partitioning of items.

Results of item-level EFA.

English. Four item-level EFA models (1-, 2-, 3-, and 4-factor models) were examined for datasets from English. Table 7 provides results pertinent to the overall model fit. The "Single" p-value is associated with the model chi-square statistic that indicates whether a particular EFA model adequately fits the observed data. The "Diff" p-value is associated with the chi-square difference statistic that evaluates whether or not adding one more factor into the existing model significantly improves the fit. A significance level of 0.05 was used in testing these chi-square statistics. In theory, an EFA model is favorable if (a) the "Single" p-value is greater than 0.05, and (b) the "Diff" p-value is smaller than 0.05. Despite the appeal of such a simple rule for model selection, model chi-square statistics seemed to provide more misleading than constructive suggestions for the best model to choose, especially as large samples were involved in the analysis. More specifically, both the "Single" and "Diff" p-values dropped below 0.05 for datasets of N = 3,000 from the 2011 and 2012 forms of English. This implies that adding one

factor to the existing model was preferred, but in the meantime, the new model always failed to fit the data well. Similar results were also found in data from Spanish and Chemistry.

The values of both RMSEA and RMSR for each dataset from English, also provided in Table 7, showed a continual decrease from the 1-factor model to the 4-factor model, suggesting that the model with more factors always fits the data better. The value of RMSEA for the four models for each dataset fell below 0.05, indicating that all the solutions fit the data reasonably well. Among them, the most parsimonious 1-factor model might be preferred. The values of RMSR only supported the use of the 1-factor model for the dataset from the 2012 form, as all RMSR values were below 0.05. For datasets from the 2011 form, the RMSR values were greater than 0.05 for any of the models examined when the sample contained 500 observations, suggesting that none of these models led to acceptable correlation residuals overall. When the sample size rose to 1,000 and 3,000, the mean absolute correlation residual was satisfactory only if the model contained two or more factors.

Numbers of substantial factor pattern coefficients and factor correlations are summarized in Tables 8 and 9. Inspection of these tables revealed that a 4-factor model was somewhat redundant. In the 4-factor model, one of the factors was not substantially related to a sufficient number of items, and sometimes, negative correlations were found between factors, which is often a sign of overfitting.

Analyses of correlation residuals partly confirmed previous findings and provided further insight into model selection. For each dataset there were a few residuals greater than 0.10 in magnitude for the four EFA models investigated. But as the number of factors increased, more and more residuals fell between -0.10 and 0.10, which might be considered to be small. For datasets from the 2011 form, a noticeable increment in proportion of small residuals was found when a 2-factor model replaced the 1-factor model. However, for the dataset from the 2012 form, such an increment could be seen but was not substantial. An examination of the pattern of residuals also revealed two special patterns for the 1-factor model for the 2011 form: (a) the correlation among MC items from the last testlet tended to be underestimated, whereas the correlations between these items and the rest of items tended to be overestimated; (b) the correlations among the three FR items were underestimated. Only after two more factors were added to the 1-factor model, all the residuals were considerably small in magnitude and these special patterns became largely weakened. Similar patterns of residuals can also be found in the

data from the 2012 form but were less substantial. This finding again drew attention to the potential difference in dimensional structure between the two forms. In addition, correlation residuals tended to have large absolute values for small samples.

Results of R-squares and R-square improvements are plotted for each dataset from English in Figure 3. The observed variances for the last several MC items and three FR items were better explained after three factors were included in the model, which provides more evidence towards a 3-factor model. A fourth factor seemed to contribute to the explanation of variance for a few MC items for the dataset of N = 500, but this factor gradually disappeared as the dataset contained over 1,000 observations. This finding leads to cautions that some of the factors uncovered by using smaller samples might only reflect estimation "errors" in EFA.

To summarize, unidimensionality might be roughly assumed for data from English, but some degree of multidimensionality was detected using item-level EFA. A 3-factor model might be more useful for understanding the exam's internal structure, especially for the 2011 form. Dimensional structure for the 2012 form differed slightly and seemed to be more nearly unidimensional. In Table 10, a simplified 3-factor solution under the baseline condition is provided, which only contains substantial pattern coefficients. Based on the 3-factor model, items on the form were grouped into three clusters according to their highest factor loadings, as shown in Table 11, where items with substantial pattern coefficients are emphasized.

Spanish. Six item-level EFA models (1-, 2-, 3-, 4-, 5-, and 6-factor models) were examined for datasets from Spanish. Table 12 shows the p-values for model chi-squares and the values of RMSEA and RMSR for the six models for each dataset from Spanish. The sensitivity of chi-squares to sample size can be seen again. The chi-square tests for the single model fit and model comparison were significant at the level of 0.05 for all six models when the sample size reached 3,000, providing little information concerning which EFA model should be selected.

For each dataset, the values of RMSEA for models with one to six factors were all smaller than 0.05, indicating acceptable model fit. Nevertheless, RMSEA was reduced almost by half when a 2-factor model took the place of the 1-factor model; RMSEA continued to decrease when more factors were included in the model, but those declines were less remarkable. Similar to that for English, the values of RMSR exceeded 0.05 for all the models examined using a small sample of N = 500 for Spanish. When the sample consisted of an adequate number of observations (e.g., N = 3,000), the RMSR criterion of 0.05 performed as expected. The value of

RMSR for the 1-factor model was greater than 0.05, again implying that the use of a single factor might be insufficient for predicting correlations between item scores.

In Tables 13 and 14, factor pattern and factor structure coefficients as well as correlations between factors are reported for each dataset from Spanish. As seen in these tables, models having five or six factors might not be interpretable; for instance, negative factor correlations occurred in the 5- and 6-factor models.

It was evident in the analyses of correlation residuals that, under the baseline condition (i.e., 2011 Form, N = 3,000), one factor alone could not accurately predict item score correlations, and adding one additional factor led to nearly 10% increment in the number of residuals between -0.10 and 0.10. This proportion of small residuals increased slightly as the third and fourth factors were added. Under the baseline condition, the following two special patterns became obscure only after four factors were used: (a) underestimation of the correlations within several MC testlets, and (b) overestimation between items from the last two testlets. Similar results can also be found in the data from the 2012 form. The use of smaller samples resulted in more extreme residuals but did not significantly affect special patterns.

Examination of R-squares and R-square improvements in Figure 4 also yields slight preference for a 4-factor model. Since a model with either five or six factors seemed least favorable, the figure only depicts R-squares and R-square improvements for models with at most four factors. Compared with using one or two factor, using at least three factors seemed to explain the observed variances of the majority of item scores for Spanish and especially improved the fit of scores on the last two testlets.

Combining the EFA results together for the Spanish datasets, multidimensionality was suggested for scores collected from Spanish, and a 4-factor model might be favorable. Table 15 provides a simplified 4-factor solution identifying only the substantial factor pattern coefficients under the baseline condition. Based on the 4-factor model, items on the test were divided into four clusters, as specified in Table 16.

Chemistry. Four item-level EFA models (1-, 2-, 3-, and 4-factor models) were examined for datasets from Chemistry. As shown in Table 17, the p-values for chi-squares again provided mixed suggestions for model selection, whereas the RMSEA and RMSR statistics tended to be more informative and robust. Compared with the other two language exams, the values of RMSEA and RMSR were generally small for the four EFA models examined for Chemistry, and

reductions in these statistics by using more complex models were not obvious. In other words, the simplest 1-factor model might be good enough to describe the observed data from Chemistry.

The factor pattern and factor structure coefficients in Table 18 as well as the factor correlations in Table 19 suggest overfitting by the 4-factor model. Furthermore, the number of items that have substantial relationships with factors steadily decreased as the number of factors increased, indicating that the use of additional factors would not improve the fit but instead complicates interpretations of the factor solution.

Analyses of correlation residuals showed that most residuals under the four EFA models were small in magnitude. Although a few extreme residuals were found when small samples were used, as the sample size was as large as 3,000, over 95% of the residuals fell between -0.10 and 0.10 under the 1-factor model. No salient patterns related to the 1-factor model were found, providing more evidence that unidimensionality might hold for the data from Chemistry.

Last, R-squares and R-square improvements for the four EFA models for the Chemistry datasets are displayed in Figure 5. The observed variances of most of the items seemed to be adequately estimated using at most two factors.

In short, essential unidimensionality can likely be assumed for data from Chemistry. A 2-factor model slightly improved estimation of variances of individual item scores and correlations between item scores, although model complexity and further issues related to multidimensional models should also be taken into consideration.

Results of MIRT cluster analysis.

English. The four-dimensional M3PL and MGR models were fit to the MC and FR item scores on the two forms of English. Irregular items listed in Tables 5 and 6 were eliminated from calibration to avoid convergence problems. For example, Item 30 was not calibrated when datasets of N = 500, 1,000, and 3,000 from the English 2011 form was used. This item was also eliminated from subsequent cluster analysis and did not receive any cluster membership. Based on the slope parameter estimates, angular distances were computed in every possible item pair, except for the pairs involving items eliminated from calibration.

Table 20 contains values of five indices calculated based on angular distances. The best number of clusters indicated by each index for each dataset from English is emphasized in bold. Given definitions of these indices, the CH, C-, and Gamma indices cannot be used to assess the unidimensionality assumption, whereas the DH and Beale indices can be. Examining this table

partly confirmed the relationship among these indices found in Milligan and Cooper (1985). In particular, the DH and Beale indices tended to indicate fewer clusters, whereas the Gamma index tended to indicate more clusters. The CH and C- indices seemed to provide similar results for the 2011 form but not for the 2012 form. Inconsistencies were also found in results pertinent to different forms and sample sizes. In short, as the true cluster structure of the real data was unknown in this study, large variability of index values made the final decision of the number of clusters become extremely difficult. To resolve this problem, this study used the optimal number of clusters indicated by the CH index to cluster items, because it was shown to best recover the true cluster structure in Milligan and Cooper (1985). Results of the CH index from the largest samples (N = 3,000) were considered.

According to the CH index, a 2-cluster solution was produced by complete linkage cluster analysis for the 2011 form, and a 4-cluster solution was produced for the 2012 form, as provided in Table 21. In addition to the columns that store items within clusters, another column named “Undecided” stores items eliminated from both MIRT calibration and cluster analysis due to peculiar item difficulty and/or discrimination. The four clusters for the 2012 form might be collapsed into fewer clusters, because two clusters — Clusters 2 and 3 — contained no more than four items. Cluster 2 seemed more interpretable as a format factor, but further information is still needed to make any definitive conclusions. A cluster dendrogram for the 2-, 3-, and 4-cluster solutions under the baseline condition can be found in Figure 6. Given the test length, test configuration, and cluster complexity in this study, it was difficult to decide at which height or level the dendrogram should be cut. Nevertheless, this figure uncovers an interesting finding: in a 4-cluster solution, all three FR items (Items 55, 56, and 57) composed a cluster independent from the MC items, which was not found in the data from Spanish and Chemistry.

Spanish. The six-dimensional M3PL and MGR models were fit to the MC and FR item scores on the two forms of Spanish. Table 22 contains the values of the CH, DH, C-, Gamma, and Beale indices regarding possible cluster solutions for each dataset from Spanish. As seen in the table, these indices exhibited the same tendency found previously in the English datasets. The DH and Beale indices indicated that the 1-cluster solution was the best solution, while unidimensionality seemed less plausible. The Gamma index indicated as many as six dimensions, which might also be due to overestimation. Results using the CH index were in the middle, ranging from two to four dimensions. Results using the C-index were similar to those

using the CH index when smaller samples were used but were more consistent with those using the Gamma index as the sample size reached 3,000.

Based on the largest samples, for the CH index a 4-cluster solution was preferred for the 2011 form and a 2-cluster solution for the 2012 form. These solutions are presented in Table 23. The cluster dendrogram in Figure 7 did not help clarify the best number of clusters but did yield more detailed information about the hierarchical structure detected between items. For example, two FR items, Items 71 and 72, were grouped together at the lowest level and then integrated with another FR item, Item 74, at the next lowest level. This might provide some evidence of close relationships among these FR items. By contrast, the other FR item, Item 73, seemed to have closer relationships with some MC items than it did with the other three FR items. This pattern could also have been found in Table 23 that shows item partition based on the 2-cluster solution for each dataset from Spanish — Item 73 was sometimes assigned to the same cluster with Item 71, 72, and 74, but at other times it was assigned to a different cluster from the other FR items.

Chemistry. The four-dimensional M3PL and MGR models were fit to the MC and FR item scores on the two forms of Chemistry. Table 24 presents the values of different criteria that were used to determine the best number of clusters in this study. Highly diverse values among the CH, DH, C-, Gamma, and Beale indices can also be found in the Chemistry datasets, which complicated the decision as to when to stop clustering. The relationship between these indices was similar to those uncovered in the data from English and Spanish, except for the CH index. For Chemistry, the CH index typically indicated three or four dimensions, while the data from Chemistry appeared to be more unidimensional. In fact, item partitioning based on the CH index partly supported overestimation. As seen in Table 25, one cluster always contained an overwhelming number of items, and another cluster contained a few items, whereas the remaining one or two clusters contained no more than four items. Except for those in the dominant cluster (Cluster 2), items in the other clusters also varied dramatically, making the resulting partition less interpretable. In Figure 8, the cluster dendrogram with the 2-, 3-, and 4-cluster solutions are shown. At lower levels, the FR items were mixed with the MC items and did not form independent clusters. Similar to those found for English and Spanish, based on the cluster dendrogram alone it was difficult to determine the number of clusters.

Comparison between item-level EFA and MIRT cluster analysis. In this study, both item-level EFA and MIRT cluster analysis provided dimensionality-based partitions of items for English, Spanish, and Chemistry. Table 26 presents the values of several statistics that represent the level of consistency between dimensional structure estimated by the two methods.

First, there was a difference in the number of items actually used in the procedure of dimensionality assessment. Item-level EFA included all the items into the estimation of factor models and assigned cluster membership to every item on the exam. When applying this method, an item might have substantial relationships with no factors or multiple factors. These issues were almost resolved as the items were grouped into clusters according to their highest factor loadings, although the resulting clusters should still be interpreted carefully. On the contrary, MIRT cluster analysis removed those with peculiar item difficulty and discrimination from the dimensionality assessment, thereby not assigning them any clusters. In this way, variations in clustering due to irregular items were avoided, though the interpretations of those "undecided" items still required consideration.

Second, discrepancies were found in the number of clusters of dimensions indicated by the two methods. Based on a comprehensive collection of criteria and guidelines, item-level EFA proposed three dimensions for data from English, four dimensions for data from Spanish, and one dimension (i.e., unidimensionality) for data from Chemistry. MIRT cluster analysis, however, only agreed with item-level EFA on the data from the Spanish 2011 form. Large discrepancies were found in the Chemistry datasets. Although item-level EFA supported the unidimensionality assumption, MIRT cluster analysis suggested three or four dimensions. When using MIRT cluster analysis, such decisions were made primarily relying on the stopping rules, since cluster dendrograms provided little help in determining the number of clusters given the complexity of data and models in this study. Inconsistencies were often found among different stopping rules, making the final decision of the number of dimensions extremely difficult.

Compared with those indicated by item-level EFA, numbers of dimensions decided using MIRT cluster analysis appeared less convincing. There is no literature clarifying the performance of stopping rules in the angular distance space. Even assuming the traditional Euclidean space, the use of different stopping rules could lead to substantially different clustering solutions. More research is still needed to evaluate their performance under realistic conditions. In addition, the assumption of unidimensionality cannot always be examined by MIRT cluster analysis,

depending on which stopping rules are followed in the clustering procedure. This feature could compromise applications of this method to situations where examination of unidimensionality is of primary concern. By contrast, criteria and guidelines used in item-level EFA have been discussed more extensively in previous studies, and some good practices have also been demonstrated (e.g., Kline, 2010; Stone & Yeh, 2006).

Third, specific item partitions varied between item-level EFA and MIRT cluster analysis, which could be reflected by an overall consistency index, φ, defined in this study. The higher the value, the more item pairs were assigned consistently to the same or different dimensions by the two methods. In Table 26, the "Original" φ was calculated using partitions originally produced by the two methods. For instance, the "Original" φ of 0.66 for the dataset of N = 500 from the English 2011 form was obtained by comparing the two clusters specified in the first row of Tables 11 and 21. Obviously, this value was affected by two sources of differences in dimensional structure: number of dimensions and ways to cluster items. It was suspected that MIRT cluster analysis might be more likely to be disturbed by estimation of the number of dimensions. Thus, φ was calculated alternatively after adjusting the number of dimensions used in MIRT cluster analysis to equal that proposed by item-level EFA. That is, for the same dataset of N = 500 from the English 2011 form, the "Adjusted" φ was obtained by comparing the following two clusters: the cluster specified in the first row of Table 11 and the cluster based on the 3-cluster solution produced by MIRT cluster analysis. Noticeable improvements in the overall consistency can be found after adjusting for the number of dimensions, and a more substantial improvement (approximately 10% increase) was shown in the dataset of N = 3,000 from the Spanish 2012 form, where the original number of dimensions estimated by the two methods differed by 2. An exception was also seen in the dataset of N = 3,000 from the English 2012 form, where the overall consistency dropped slightly after adjusting for the number of dimensions. In general, item-level EFA and MIRT cluster analysis tended to perform similarly once numbers of dimensions were set equal. For data from Chemistry, no "Adjusted" φ was available due to lack of examination of unidimensionality of MIRT cluster analysis.

Sources of Multidimensionality

Tables 27 and 28 show the dimensional solutions produced by item-level EFA and MIRT cluster analysis for English and Spanish, respectively. For each exam, only the baseline condition

(2011 form, N = 3,000) was considered. No such table was created for Chemistry because unidimensionality likely held for this score, and as such no clusters were formed.

In Table 27 for English, item-level EFA suggested the following relationships between items and three dimensions: (a) MC items from the first reading testlet (R1) and three FR writing items were primarily related to one dimension (Dimension 1); (b) MC items from the third reading testlet (R3) were related to another dimension (Dimension 2); (c) MC items from the second reading testlet (R2) tapped the above two dimensions (Dimensions 1 and 2); and (d) MC items from the fourth reading testlet (R4) tended to measure a third dimension (Dimension 3). MIRT cluster analysis estimated the dimensional structure in a rougher manner, as it combined the first two dimensions estimated using item-level EFA into one dimension. Since the FR items were always grouped with some of the MC items, these dimensional solutions did not reflect an apparent format effect; though, as indicated previously, the FR items formed an independent cluster at a lower level of the hierarchical structure among items. Instead, the content distribution along with the testlet dependence might play a major role. One dimension detected by either method (Dimension 3 by item-level EFA, and equivalently, Dimension 2 by MIRT cluster analysis) was clearly associated with one of the content domains on the English exam. Within this content domain, item-level EFA further separated two dimensions based on the testlets, whereas MIRT cluster analysis did not.

Table 28 reveals a more complicated dimensional structure for Spanish. Again it was uncertain whether the format effects were present, because the FR items did not form a dimension distinguished from the MC items. Evidence concerning the content effects was also obscure given the fact that some dimensions were related with items from different content domains. Dimensions 3 and 4 were represented by two reading testlets, R4 and R3, respectively. This might be due to some unique content measured by these testlets, or perhaps by other test characteristics shared by the items nested within these testlets. Further documentation and analyses would be needed to justify plausible explanations.

Discussion

The purpose of this study was to empirically compare two promising dimensionality assessment methods for exploring the dimensional structure of mixed-format tests. Data from three AP exams were used as illustrative examples. Specific methods applied included: item-level EFA and MIRT cluster analysis, which rely on different definitions of dimensionality (i.e.,

EFA and IRT perspectives) and use different procedures for clustering items. A significant distinction between the two methods is the direction of examination of possible dimensional structure for the given data. Item-level EFA begins with the simplest 1-factor model, namely the unidimensional model. If the 1-factor model appears untenable, this method adds one factor at a time to the existing model and decides whether the new model better fits the data. MIRT cluster analysis, however, starts with a sufficiently large number of dimensions. Such an overly high dimensional space further collapses into a lower dimensional space by checking the similarities among items. This study is the first to compare the two methods in terms of dimensional structure and item partitioning. Inconsistencies between the two methods included the number of items actually used in dimensionality assessment, the number of dimensions decided, and the clustering solution. Once the number of dimensions was set to equal between the two methods, the resulting clustering solutions tended to be more similar. This study also found substantial inconsistencies among multiple stopping criteria for each method, especially for MIRT cluster analysis. Since the relevant literature is scarce for this method, the best stopping criterion to follow remains uncertain.

Data from three mixed-format exams, English, Spanish, and Chemistry, were used in this study. Although certain discrepancies occurred between item-level EFA and MIRT cluster analysis, data from two language exams, especially those from Spanish, tended to be multidimensional. In contrast, data from Chemistry tended to be essentially unidimensional. These findings are consistent with many previous studies. In those studies, data from the language tests also showed more complex dimensional structure when compared to the math and science tests (e.g., Lissitz et al., 2012; Perkhounkova & Dunbar, 1999).

In large-scale testing, multiple forms are typically constructed to the same or similar test specifications, and these forms are administered at different times and to different groups of examinees. It has been suggested that dimensionality be checked for parallel forms with different samples (e.g., Reckase, 2009; Stone & Yeh, 2006). In this study, the 2011 and 2012 forms of each exam were considered, where the 2011 form was an old form and the 2012 form was a new form equated back to the 2011 form under the CINEG design. Examinees taking these forms were assumed to be nonequivalent, but the difference in their abilities was small as reflected by their scores on the common items. Performance of dimensionality assessment methods used in this study was generally consistent across parallel forms of each exam.

Sample size also played a subtle role in estimation of dimensional structure. For item-level EFA, a small sample (e.g., 500) was less favorable because it led to inaccurate estimation of factor solutions and substantial correlation residuals. However, a large sample (e.g., 3,000) also made some stopping criteria less effective, such as the chi-square tests. Several inconsistencies can be found among stopping criteria when samples of different sizes were used for MIRT cluster analysis, and more research is necessary to figure out the effect of sample size on these criteria and final item clusters.

In this study, item-level EFA and MIRT cluster analysis were further used to investigate potential sources of multidimensionality, especially the format effects. Item-level EFA and MIRT cluster analysis did not provide clear evidence concerning the format effects for the data from English and Spanish, as the FR items did not form clusters separate from the MC items at higher levels in a hierarchy. Instead, item clusters found using these methods implied that the content distribution and testlet dependence might be associated with dimensional structure of scores. Data from Chemistry were not checked because unidimensionality likely held. An alternative might be to consider some confirmatory dimensionality assessment methods, for example, model selection in CFA. A primary reason is that, when researchers and practitioners scrutinize dimensions caused by different item types or other sources, they are likely to hypothesize some dimensional structure for the data and have already collected supporting evidence. Under this circumstance, confirmatory dimensionality assessment might be more reasonable and powerful than exploratory dimensionality assessment discussed in this study.

The scope of this study was limited to the use of two promising methods for exploring dimensional structure for mixed-format tests. The item-level EFA and MIRT cluster analysis methods were selected for this study because these methods are well-established, ready to implement, and represent the two widely adopted perspectives towards dimensionality, namely the perspectives from factor analysis and IRT. In order to expand the measurement practitioner's toolbox, future research should investigate application of other possible dimensionality assessment methods for mixed-format tests. For example, some nonparametric methods that are more data-oriented and require few models and assumptions could be considered. In some situations (e.g., the correlations between dimensions are high), because of their nonparametric feature, those methods might show outstanding performance. Different data sources should also be examined in the future research. Using real data from three large-scale mixed-format tests in

this study allowed for potential similarities and dissimilarities among methods to be observed under realistic conditions. Using simulated data in the future research would help determine what factors contribute to the similarities and dissimilarities among methods. These factors are usually hard to control or manipulate in a real data study, such as the number of dimensions, the correlation between dimensions, and certain types of test and sample irregularities. Pseudo-test forms and pseudo-groups might also be useful when comparing different dimensionality assessment methods, as they preserve some of the test and sample characteristics of the real data and control for the rest. Finally, although test items, especially the MC and FR items, often receive different weights in calculating a composite score, they were treated as being equal in the process of dimensionality assessment in this study. More research is still needed to fully understand the possible consequences of ignoring the test item weighting.

References

Abdi, H. (2003). Rotations. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of Social Sciences Research Methods* (pp. 979–983). Thousand Oaks, CA: SAGE Publications, Inc.

Baker, F. B., & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association, 70*, 31–38.

Beale, E. M. L. (1969). *Cluster analysis*. London: Scientific Control Systems.

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

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*, 313–329.

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

Caliński, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1–27.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*, 1–36.

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

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Hattie, J. A. (1981). *Decision criteria for determining unidimensionality* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global.

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

Hohensin, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732–746.

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83*, 1072–1080.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*, 1–21.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56*, 255–278.

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

Kolen, M. J., & Lee, W. (2014). Introduction and overview. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3).* (CASMA Monograph Number 2.3). Iowa City, IA: CASMA, The University of Iowa.

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*, 1–50.

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education, 5*, 193–211.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education, 19*, 67–91.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73–90.

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

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99–117.

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

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.

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

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 73–103.

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.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

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

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment, 19*, 35–57.

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., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*, 355–369.

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.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.

Zhang, M., Kolen, M. J., & Lee, W. (2014). A comparison of test dimensionality assessment approaches 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 Number 2.3). Iowa City, IA: CASMA, The University of Iowa.

Zwick, R. J. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293–308.

Table 1

Item and Total Score Points for the Selected Exams

Exam	Form	MC Section	FR Section	MC Total	FR Total	Total
English	2011	54 (1 each)	3 (9 each)	54	27	81
	2012	55 (1 each)	3 (9 each)	55	27	82
Spanish	2011	70 (1 each)	4 (5 each)	70	20	90
	2012	70 (1 each)	4 (5 each)	70	20	90
Chemistry	2011	75 (1 each)	6 (9, 9, 9, 7, 8, 8)	75	50	125
	2012	75 (1 each)	6 (9, 9, 9, 7, 8, 9)	75	51	126

Table 2

Calibration Settings for MIRT Calibration in flexMIRT

Setting	Value
Number of cycles for E steps	1,500
Number of cycles for M steps	1,500
Convergence criterion for E steps	0.0001
Convergence criterion for M steps	0.0001
Number of quadrature points	21 (for ≤ 4 dimensions) or 7 (for ≥ 5 dimensions)
Quadrature range	[-5, 5]
Slope prior	Normal (1.0, 1.0)
Location prior	None
Pseudo-guessing prior	Beta (2,5) (for English and Chemistry, 4-option MC) Beta (2,4) (for Spanish, 5-option MC)

Table 3

Total Score Range and Moments

Exam	Form	Range	*N*	Mean	SD	Skew	Kurt
English	2011	0–81	500	48.24	12.72	-0.34	2.55
			1,000	48.84	12.13	-0.36	2.66
			3,000	48.73	12.73	-0.53	2.89
	2012	0–82	3,000	47.34	12.19	-0.18	2.65
Spanish	2011	0–90	500	57.46	13.94	-0.41	2.93
			1,000	58.08	13.25	-0.40	2.90
			3,000	57.70	13.76	-0.40	2.88
	2012	0–90	3,000	61.03	13.28	-0.52	3.16
Chemistry	2011	0–125	500	63.13	23.37	0.02	2.27
			1,000	61.51	24.82	0.08	2.20
			3,000	60.62	24.57	0.11	2.25
	2012	0–126	3,000	61.18	23.26	0.09	2.32

Table 4

Reliability Coefficients and Disattenuated MC and FR Correlations

			Reliability			Disattenuated
Exam	Form	*N*	MC	CR	Total	MC-FR Correlation
English	2011	500	0.90	0.72	0.91	0.79
		1,000	0.89	0.69	0.90	0.78
		3,000	0.90	0.71	0.91	0.81
	2012	3,000	0.87	0.69	0.89	0.81
Spanish	2011	500	0.91	0.65	0.92	0.84
		1,000	0.90	0.64	0.91	0.82
		3,000	0.90	0.70	0.92	0.82
	2012	3,000	0.91	0.68	0.92	0.82
Chemistry	2011	500	0.92	0.88	0.95	0.97
		1,000	0.93	0.89	0.96	0.98
		3,000	0.93	0.89	0.96	0.98
	2012	3,000	0.92	0.89	0.95	0.96

Note. Reliability coefficients using Cronbach's alpha were estimated separately for MC and FR sections, and reliability of total scores was calculated using stratified alpha where strata represent item format.

Table 5

Item Difficulty, Discrimination, and Irregular Items for MC Items

Exam	Form	*N*	Difficulty		Discrimination		Irregular Item
			Mean	SD	Mean	SD	No.
English	2011	500	0.64	0.16	0.50	0.16	30
		1,000	0.65	0.16	0.47	0.17	30
		3,000	0.65	0.16	0.50	0.16	30
	2012	3,000	0.61	0.17	0.43	0.13	None
Spanish	2011	500	0.64	0.17	0.46	0.14	None
		1,000	0.65	0.17	0.44	0.13	16
		3,000	0.65	0.17	0.45	0.12	16
	2012	3,000	0.68	0.17	0.45	0.14	11, 27, 44
Chemistry	2011	500	0.57	0.18	0.50	0.15	62
		1,000	0.56	0.17	0.53	0.15	None
		3,000	0.55	0.17	0.53	0.15	None
	2012	3,000	0.52	0.17	0.48	0.15	65

Table 6

Item Difficulty, Discrimination, and Irregular Items for FR Items

Exam	Form	*N*	Difficulty		Discrimination		Irregular Item
			Mean	SD	Mean	SD	No.
English	2011	500	0.50	0.02	0.57	0.01	None
		1,000	0.51	0.03	0.54	0.02	None
		3,000	0.51	0.03	0.57	0.01	None
	2012	3,000	0.52	0.03	0.55	0.04	None
Spanish	2011	500	0.62	0.07	0.51	0.03	None
		1,000	0.62	0.07	0.50	0.05	None
		3,000	0.62	0.06	0.54	0.03	None
	2012	3,000	0.66	0.08	0.53	0.02	None
Chemistry	2011	500	0.40	0.03	0.75	0.03	None
		1,000	0.39	0.03	0.77	0.04	None
		3,000	0.38	0.03	0.77	0.04	None
	2012	3,000	0.44	0.07	0.75	0.04	None

Table 7

Values of Model-Fit Statistics for the English Datasets

Form	*N*	Model	Model Chi-Square		RMSEA	RMSR
			Single	Diff		
2011	500	1-Factor	0.000*		0.024	0.078
		2-Factor	0.101	0.000*	0.010	0.062
		3-Factor	0.294	0.000*	0.006	0.059
		4-Factor	0.431	0.025*	0.004	0.056
	1,000	1-Factor	0.000*		0.029	0.069
		2-Factor	0.000*	0.000*	0.013	0.048
		3-Factor	0.002*	0.000*	0.010	0.046
		4-Factor	0.107	0.000*	0.007	0.043
	3,000	1-Factor	0.000*		0.032	0.054
		2-Factor	0.000*	0.000*	0.015	0.032
		3-Factor	0.000*	0.000*	0.012	0.030
		4-Factor	0.000*	0.000*	0.010	0.027
2012	3,000	1-Factor	0.000*		0.019	0.041
		2-Factor	0.000*	0.000*	0.014	0.033
		3-Factor	0.000*	0.000*	0.011	0.031
		4-Factor	0.000*	0.000*	0.009	0.029

Note. Baseline results are highlighted. * $p < 0.05$.

Table 8

Number of Items Substantially Related to Factors for the English Datasets

Form	*N*	1-Factor	2-Factor		3-Factor			4-Factor			
		1	1	2	1	2	3	1	2	3	4
2011	500	50	38	11	12	26	11	12	27	11	5
	1,000	50	40	13	20	15	12	23	14	11	0
	3,000	50	43	12	20	17	12	19	17	3	12
2012	3,000	53	40	9	9	15	8	16	3	10	9

Note. Baseline results are highlighted.

Table 9

Factor Correlations for the English Datasets

			2-Factor		3-Factor			4-Factor			
Form	*N*		1	2	1	2	3	1	2	3	4
		2	0.57		0.67			0.65			
	500	3			0.45	0.50		0.44	0.48		
		4						0.17	0.17	0.09	
		2	0.57		0.72			0.64			
2011	1,000	3			0.51	0.53		0.51	0.49		
		4						-0.01	-0.24	-0.01	
		2	0.60		0.65			0.69			
	3,000	3			0.51	0.52		0.19	0.30		
		4						0.49	0.58	0.25	
		2	0.63		0.75			0.46			
2012	3,000	3			0.54	0.50		0.71	0.47		
		4						0.58	0.38	0.60	

Note. Baseline results are highlighted.

Table 10

A 3-Factor Solution for the English Exam, Baseline (2011 Form, N = 3,000)

	3-Factor Solution				3-Factor Solution		
Item	1	2	3	Item	1	2	3
56	0.71			33	0.31	0.37	
57	0.69			26		0.36	
55	0.69			39		0.35	
5	0.50	0.31		37		0.35	
3	0.50			40	0.31	0.34	
13	0.43			7	0.30	0.33	
2	0.42			45		0.30	
11	0.41			8		—	
12	0.40			15		—	
1	0.40			41		—	
28	0.39			10		—	
6	0.38	0.32		20		—	
34	0.37	0.34		38		—	
17	0.34			19		—	
9	0.33			27		—	
4	0.33			30		—	
14	0.31			53			0.98
23	—			52			0.93
25	—			54			0.79
22	—			50			0.66
24	—			48			0.54
18	—			51			0.51
16	—			49			0.42
36		0.48		47			0.40
31		0.44		44			0.38
35		0.42		42			0.37
32		0.41		43			0.33
21		0.39		46		0.32	0.32
29		0.39					

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

Table 11

Item Clusters Produced by Item-Level EFA for the English Datasets

Form	*N*	Cluster 1	Cluster 2	Cluster 3
2011	500	**1 2 3 7 9** 12 **13** 14 18 **22** 23 **28 55 56 57**	**4 5 6 8** 10 **11** 15 16 **17 19** 20 21 24 **25 26** 27 **29** 30 **31 32 33 34 35 36 37 38 39 40 41 42 43 45**	**44 46 47 48 49 50 51 52 53 54**
	1,000	**1 2 3 4 6 7 8 9 11** 12 **13** 14 16 **17 19 20** 21 **22** 24 **25** 38 **55 56 57**	**5** 10 **15** 18 23 **26** 27 **28 29** 30 **31** 32 **33 34 35 36 37 39 40 41**	**42 43 44** 45 **46 47 48 49 50 51 52 53 54**
	3,000	**1 2 3 4 5 6 9 11 12 13 14** 16 **17** 18 22 23 24 25 **28 34 55 56 57**	**7** 8 10 15 19 20 **21 26** 27 **29** 30 **31 32 33 35 36 37** 38 **39 40** 41 **45**	**42 43 44 46 47 48 49 50 51 52 53 54**
2012	3,000	**1 2 3 4 6 7 8** 9 10 11 12 13 **14** 15 16 **17** 18 20 21 **27** 28 29 **32 34 56 57 58**	5 19 22 23 24 **25** 26 30 31 33 **35 36 37 38 39 40** 41 **42 43** 44	45 **46** 47 **48 49 50** 51 **52 53 54 55**

Note. Baseline results are highlighted. Items substantially related to the factor are emphasized in bold. The order of clusters is not necessarily the order of factors outputted by Mplus.

Table 12

Values of Model-Fit Statistics for the Spanish Datasets

Form	*N*	Model	Model Chi-Square		RMSEA	RMSR
			Single	Diff		
2011	500	1-Factor	0.000*		0.027	0.092
		2-Factor	0.000*	0.000*	0.014	0.071
		3-Factor	0.024*	0.000*	0.011	0.066
		4-Factor	0.120	0.000*	0.008	0.063
		5-Factor	0.231	0.003*	0.007	0.059
		6-Factor	0.380	0.005*	0.004	0.057
	1,000	1-Factor	0.000*		0.028	0.077
		2-Factor	0.000*	0.000*	0.016	0.057
		3-Factor	0.000*	0.000*	0.014	0.052
		4-Factor	0.000*	0.000*	0.011	0.048
		5-Factor	0.004*	0.000*	0.009	0.046
		6-Factor	0.017*	0.000*	0.008	0.044
	3,000	1-Factor	0.000*		0.034	0.067
		2-Factor	0.000*	0.000*	0.020	0.042
		3-Factor	0.000*	0.000*	0.016	0.036
		4-Factor	0.000*	0.000*	0.013	0.032
		5-Factor	0.000*	0.000*	0.011	0.030
		6-Factor	0.000*	0.000*	0.010	0.028
2012	3,000	1-Factor	0.000*		0.031	0.063
		2-Factor	0.000*	0.000*	0.018	0.045
		3-Factor	0.000*	0.000*	0.015	0.040
		4-Factor	0.000*	0.000*	0.013	0.035
		5-Factor	0.000*	0.000*	0.012	0.033
		6-Factor	0.000*	0.000*	0.010	0.031

Note. Baseline results are highlighted. * $p < 0.05$.

Table 13

Number of Items Substantially Related to Factors for the Spanish Datasets

		1-Factor	2-Factor		3-Factor			4-Factor			
Form	*N*	1	1	2	1	2	3	1	2	3	4
	500	50	38	11	12	26	11	12	27	11	5
2011	1,000	50	40	13	20	15	12	23	14	11	0
	3,000	50	43	12	20	17	12	19	17	3	12
2012	3,000	53	40	9	9	15	8	16	3	10	9

		5-Factor					6-Factor					
Form	*N*	1	2	3	4	5	1	2	3	4	5	6
	500	21	5	20	16	13	19	6	0	19	9	9
2011	1,000	8	14	12	24	9	10	9	12	22	7	9
	3,000	0	20	23	8	8	22	20	0	8	8	3
2012	3,000	14	9	21	9	10	14	9	20	9	0	8

Note. Baseline results are highlighted.

Table 14

Factor Correlations for the Spanish Datasets

			2-Factor		3-Factor			4-Factor			
Form	*N*		1	2	1	2	3	1	2	3	4
		2	0.55		0.59			0.49			
	500	3			0.49	0.55		0.46	0.41		
		4						0.46	0.57	0.36	
		2	0.59		0.54			0.52			
2011	1,000	3			0.59	0.46		0.53	0.32		
		4						0.47	0.43	0.37	
		2	0.60		0.56			0.57			
	3,000	3			0.55	0.46		0.48	0.30		
		4						0.54	0.49	0.42	
		2	0.63		0.59			0.40			
2012	3,000	3			0.55	0.57		0.61	0.43		
		4						0.62	0.35	0.49	

			5-Factor					6-Factor					
Form	*N*		1	2	3	4	5	1	2	3	4	5	6
		2	0.05					0.10					
		3	0.46	-0.03				-0.30	-0.18				
	500	4	0.47	0.10	0.41			0.38	0.13	-0.39			
		5	0.45	0.07	0.56	0.39		0.33	0.37	-0.34	0.26		
		6						0.48	0.16	-0.34	0.43	0.33	
		2	0.36					0.36					
		3	0.28	0.45				0.32	0.40				
2011	1,000	4	0.19	0.44	0.34			0.11	0.39	0.31			
		5	0.31	0.52	0.38	0.34		0.32	0.49	0.49	0.25		
		6						0.28	0.37	0.44	0.30	0.48	
		2	-0.47					0.50					
		3	-0.44	0.52				-0.36	-0.36				
	3,000	4	-0.40	0.53	0.56			0.55	0.50	-0.33			
		5	-0.20	0.17	0.41	0.40		0.44	0.20	-0.20	0.42		
		6						0.33	0.45	-0.10	0.26	0.18	
		2	0.49					0.53					
		3	0.48	0.54				0.59	0.51				
2012	3,000	4	0.47	0.51	0.30			0.56	0.46	0.42			
		5	0.51	0.51	0.56	0.49		-0.26	-0.08	-0.24	-0.38		
		6						0.47	0.56	0.56	0.46	-0.09	

Note. Baseline results are highlighted.

Table 15

A 3-Factor Solution for the Spanish Exam, Baseline (2011 Form, N = 3,000)

	4-Factor Solution					4-Factor Solution			
Item	1	2	3	4	Item	1	2	3	4
44	0.72				38		0.65		
49	0.68				23		0.63		
48	0.58				39		0.55		
14	0.55				37		0.53		
46	0.54				40		0.49		
51	0.52				41		0.47		
34	0.48				43		0.44		
26	0.47				36		0.42		
2	0.46				17		0.42		
33	0.45				74		0.41		
15	0.45				22		0.41		
1	0.44				4		0.37		
50	0.44				71		0.36		
42	0.43				25		0.31		
45	0.41				18		—		
30	0.41				31		—		
52	0.39				21		—		
29	0.38				28		—		
47	0.38				16		—		
6	0.37				69			0.76	
7	—				68			0.64	
9	—				64			0.59	
5	—				66			0.51	
72	—				70			0.45	
32	—				62			0.40	
3	—				65			0.39	
27	—				67			0.34	
19	—				63			0.33	
35	—				58				0.62
13		0.87			55				0.60
73		0.85			53				0.59
12		0.76			54				0.57
10		0.71			61				0.56
24		0.69			56				0.52
20		0.69			59				0.50
8		0.69			57				0.44
11		0.67			60				—

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

Table 16

Item Clusters Produced by Item-Level EFA for the Spanish Datasets

Form	*N*	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2011	500	**1 10 11 12 13** 19 **22** 27 28 **36 37 40 50 53 54 55 56 58 71 72 73 74**	**2 3 5** 6 7 **9 14 15** 16 18 **26** 29 **31 33** 34 42 **44 45 46 47 48 49 51 52**	57 **59** 60 **61 62 63 64 65 66 67 68 69 70**	**4 8** 17 **20** 21 **23 24 25** 30 **32** 35 **38 39 41 43**
	1,000	**1 2 3 4** 5 **6** 7 **9 14 15** 16 18 **26 27 29** 30 32 **33 34 42 44 45 46** 47 **48 49 50 51** 72	**8 10 11 12 13 17** 19 **20 21 22 23 24 25** 31 35 **36 37 38 39 40 41 43 71 73 74**	**62** 63 **64 65 66** 67 **68 69 70**	28 **52 53 54 55 56 57 58 59** 60 **61**
	3,000	**1 2** 3 5 **6** 7 9 **14 15** 19 **26** 27 **29 30** 32 **33 34** 35 **42 44 45 46 47 48 49 50 51 52** 72	**4 8 10 11 12 13** 16 **17** 18 **20** 21 **22 23 24 25** 28 31 **36 37 38 39 40 41 43 71 73 74**	**62 63 64 65 66 67 68 69 70**	**53 54 55 56 57 58 59** 60 **61**
2012	3,000	**1 2** 3 **4 10 11 17 19 26 27 28** 34 **44 46 48 49 50 51**	6 **7 8 12** 13 15 **20** 21 **22 23 24 25** 29 **30 31** 32 **33** 35 **36 37 38 39 40 41 43** 45 **47 54 71 73 74**	**5** 9 **14** 16 18 **42** 52 **62** 63 **64 65 66 67 68 69 70 72**	**53 55 56 57 58 59 60 61**

Note. Baseline results are highlighted. Items substantially related to the factor are emphasized in bold. The order of clusters is not necessarily the order of factors outputted by Mplus.

Table 17

Values of Model-Fit Statistics for the Chemistry Datasets

Form	*N*	Model	Model Chi-Square		RMSEA	RMSR
			Single	Diff		
2011	500	1-Factor	0.000*		0.013	0.069
		2-Factor	0.005*	0.000*	0.012	0.065
		3-Factor	0.016*	0.000*	0.011	0.062
		4-Factor	0.038*	0.001*	0.010	0.060
	1,000	1-Factor	0.000*		0.015	0.052
		2-Factor	0.000*	0.000*	0.012	0.045
		3-Factor	0.000*	0.000*	0.010	0.043
		4-Factor	0.003*	0.000*	0.009	0.041
	3,000	1-Factor	0.000*		0.019	0.040
		2-Factor	0.000*	0.000*	0.015	0.032
		3-Factor	0.000*	0.000*	0.014	0.030
		4-Factor	0.000*	0.000*	0.012	0.028
2012	3,000	1-Factor	0.000*		0.016	0.036
		2-Factor	0.000*	0.000*	0.013	0.031
		3-Factor	0.000*	0.000*	0.011	0.029
		4-Factor	0.000*	0.000*	0.010	0.027

Note. Baseline results are highlighted. * $p < 0.05$.

Table 18

Number of Items Substantially Related to Factors for the Chemistry Datasets

Form	*N*	1-Factor	2-Factor		3-Factor			4-Factor			
		1	1	2	1	2	3	1	2	3	4
2011	500	75	41	28	18	13	35	0	19	33	14
	1,000	78	35	43	23	26	26	26	6	10	27
	3,000	78	41	35	29	30	9	24	8	27	7
2012	3,000	72	33	32	30	17	13	26	18	10	2

Note. Baseline results are highlighted.

Table 19

Factor Correlations for the Chemistry Datasets

			2-Factor		3-Factor			4-Factor			
Form	*N*		1	2	1	2	3	1	2	3	4
2011	500	2	0.66		0.47			-0.28			
		3			0.61	0.53		-0.36	0.57		
		4						-0.07	0.35	0.43	
	1,000	2	0.67		0.52			0.55			
		3			0.56	0.56		0.61	0.46		
		4						0.57	0.48	0.62	
	3,000	2	0.68		0.63			0.54			
		3			0.53	0.55		0.55	0.59		
		4						0.43	0.48	0.44	
2012	3,000	2	0.68		0.65			0.64			
		3			0.49	0.62		0.48	0.62		
		4						0.44	0.55	0.46	

Note. Baseline results are highlighted.

Table 20

Stopping Rules and Suggested Number of Clusters for the English Datasets

Form	*N*	Solution	CH	DH	C-index	Gamma	Beale
2011	500	1-Cluster	—	**0.69**	—	—	**1.06**
		2-Cluster	**24.17**	0.88	0.40	0.63	0.32
		3-Cluster	15.36	0.78	**0.38**	0.71	0.65
		4-Cluster	13.72	0.68	0.41	**0.84**	0.93
	1,000	1-Cluster	—	**0.70**	—	—	**1.03**
		2-Cluster	**23.58**	0.78	0.36	0.67	0.68
		3-Cluster	15.12	0.73	**0.32**	0.69	0.81
		4-Cluster	15.88	0.69	0.39	**0.79**	1.04
	3,000	1-Cluster	—	**0.71**	—	—	**0.97**
		2-Cluster	**22.13**	0.73	**0.34**	0.70	0.84
		3-Cluster	19.61	0.74	0.36	0.77	0.84
		4-Cluster	16.08	0.86	0.37	**0.89**	0.39
2012	3,000	1-Cluster	—	**0.82**	—	—	**0.51**
		2-Cluster	11.97	0.60	**0.29**	0.78	1.49
		3-Cluster	15.03	0.45	0.30	0.85	2.69
		4-Cluster	**16.90**	0.87	0.33	**0.91**	0.34

Note. Baseline results are highlighted. Values of stopping rules associated with the best solution are emphasized in bold.

Table 21

Item Clusters Produced by MIRT Cluster Analysis for the English Datasets

Form	*N*	Cluster 1	Cluster 2	Undecided
2011	500	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 25 26 27 28 29 31 32 33 34 35 36 37 38 39 40 41 43 45 55 56 57	21 42 44 46 47 48 49 50 51 52 53 54	30
	1,000	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 31 32 33 34 35 36 37 38 39 40 41 45	42 43 44 46 47 48 49 50 51 52 53 54 55 56 57	30
	3,000	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 31 32 33 34 35 38 40 55 56 57	36 37 39 41 42 43 44 45 46 47 48 49 50 51 52 53 54	30

Form	*N*	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Undecided
2012	3,000	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 29 30 31 32 33 34 35 36 37 38 41 43 44 47	27 56 57 58	39 40 42	45 46 48 49 50 51 52 53 54 55	None

Note. Baseline results are highlighted. Undecided = items eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

Table 22

Stopping Rules and Suggested Number of Clusters for the Spanish Datasets

Form	*N*	Solution	CH	DH	C-index	Gamma	Beale
2011	500	1-Cluster	—	**0.87**	—	—	**0.55**
		2-Cluster	10.37	0.85	0.45	0.48	0.64
		3-Cluster	11.54	0.78	**0.45**	0.42	1.07
		4-Cluster	**13.22**	0.77	0.45	0.59	1.12
		5-Cluster	12.94	0.81	0.48	0.64	0.84
		6-Cluster	11.18	1.02	0.47	**0.66**	-0.08
	1,000	1-Cluster	—	**0.78**	—	—	**1.07**
		2-Cluster	**20.09**	0.81	0.48	0.35	0.88
		3-Cluster	15.91	0.58	**0.48**	0.57	2.61
		4-Cluster	14.37	1.02	0.49	0.60	-0.06
		5-Cluster	11.25	0.77	0.52	0.63	1.08
		6-Cluster	11.62	0.76	0.50	**0.65**	1.18
	3,000	1-Cluster	—	**0.78**	—	—	**1.05**
		2-Cluster	19.70	0.45	0.43	0.57	4.50
		3-Cluster	20.40	0.71	0.49	0.60	1.54
		4-Cluster	**24.53**	0.74	0.47	0.62	1.31
		5-Cluster	22.92	0.71	**0.41**	0.73	1.49
		6-Cluster	19.68	0.70	0.44	**0.78**	1.54
2012	3,000	1-Cluster	—	**0.80**	—	—	**0.94**
		2-Cluster	**17.05**	0.84	0.48	0.23	0.72
		3-Cluster	14.92	0.78	0.47	0.46	1.01
		4-Cluster	10.86	0.71	0.48	0.49	1.56
		5-Cluster	14.21	0.88	**0.45**	0.65	0.50
		6-Cluster	12.71	0.79	0.50	**0.72**	0.98

Note. Baseline results are highlighted. Values of stopping rules associated with the best solution are emphasized in bold.

Table 23

Item Clusters Produced by MIRT Cluster Analysis for the Spanish Datasets

Form	*N*	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Undecided
2011	500	1 14 15 26 44 45 46 48 49 51 52	2 3 4 5 7 19 21 24 25 27 30 31 32 36 74	6 9 16 18 29 34 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70	8 10 11 12 13 17 20 22 23 28 33 35 37 38 39 40 41 42 43 47 50 71 72 73	None
	1,000	9 14 15 30 44 45 48 49 51 67	5 6 8 10 11 12 13 17 20 21 22 23 24 25 31 32 33 34 35 36 37 38 39 40 41 43 73	62 64 65 66 68 69 70	1 2 3 4 7 18 19 26 27 28 29 42 46 47 50 52 53 54 55 56 57 58 59 60 61 63 71 72 74	16
	3,000	1 2 9 14 15 26 42 44 45 46 47 48 49 50 51 52	8 10 11 12 13 17 20 23 24 25 28 31 38 39 40 43 73	62 64 66 68 69 70	3 4 5 6 7 18 19 21 22 27 29 30 32 33 34 35 36 37 41 53 54 55 56 57 58 59 60 61 63 65 67 71 72 74	16

Form	*N*	Cluster 1	Cluster 2	Undecided
2012	3,000	1 2 3 4 5 6 9 14 15 16 17 18 19 26 28 29 34 35 36 37 40 41 42 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 74	7 8 10 12 13 20 21 22 23 24 25 30 31 32 33 38 39 43 73	11 27 44

Note. Baseline results are highlighted. Undecided = items eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

Table 24

Stopping Rules and Suggested Number of Clusters for the Chemistry Datasets

Form	*N*	Solution	CH	DH	C-index	Gamma	Beale
2011	500	1-Cluster	—	**0.87**	—	—	**0.35**
		2-Cluster	**11.40**	0.88	**0.34**	0.41	0.32
		3-Cluster	9.91	1.04	0.37	0.49	-0.09
		4-Cluster	7.99	0.87	0.41	**0.54**	0.36
	1,000	1-Cluster	—	**0.92**	—	—	**0.20**
		2-Cluster	6.59	0.42	**0.33**	0.64	2.95
		3-Cluster	6.52	0.88	0.38	0.66	0.32
		4-Cluster	**8.06**	0.88	0.42	**0.75**	0.33
	3,000	1-Cluster	—	**0.91**	—	—	**0.24**
		2-Cluster	7.83	0.89	0.37	**0.91**	0.31
		3-Cluster	9.19	0.44	**0.34**	0.44	2.97
		4-Cluster	**16.56**	0.88	0.35	0.48	0.33
2012	3,000	1-Cluster	—	**0.85**	—	—	**0.43**
		2-Cluster	13.99	0.75	**0.32**	0.44	0.78
		3-Cluster	**16.89**	0.89	0.37	0.53	0.29
		4-Cluster	12.48	0.94	0.36	**0.56**	0.14

Note. Baseline results are highlighted. Values of stopping rules associated with the best solution are emphasized in bold.

Table 25

Item Clusters Produced by MIRT Cluster Analysis for the Chemistry Datasets

Form	*N*	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Undecided
2011	500	1	2 3 6 7 9 10 11 12 19 25 27 28 29 30 31 32 33 34 35 37 38 39 40 41 43 44 45 50 52 54 55 58 59 60 61 69 71 73 74 76 78 79 80 81	4 5 8 17 21 22 23 24 47 48 49 53 57 64 65 77	13 14 15 16 18 20 26 36 42 46 51 56 63 66 67 68 70 72 75	62
	1,000	1 2 3	4 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 36 37 38 39 40 41 42 43 44 45 47 49 52 54 55 56 58 59 60 61 63 64 65 66 67 68 69 70 71 72 76 77 78 79 80 81	5 35 46 48 51 53 57 62	6 7 50 73 74 75	None
	3,000	1 2 3 10 11 12 13 17 19 20 22 28 29 30 31 38 44 47 49 59 76	4 5 7 8 9 15 18 21 23 24 25 26 27 32 33 34 35 36 37 39 40 41 42 43 45 46 48 50 51 52 53 54 55 56 57 58 60 61 63 64 65 66 67 68 69 70 71 72 73 77 78 79 80 81	6 62 74 75	14 16	None

Note. Baseline results are highlighted. Undecided = items eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

Table 25 *(Cont.)*

Form	*N*	Cluster 1	Cluster 2	Cluster 3	Undecided
2012	3,000	1 2 3 6 7 8 10 12 13 28 29 34 37 38 41 42 51 52 53 54 55 56 60 61 67 74 75 76 78 79 80 81	4 5 9 11 14 15 16 18 20 21 22 23 24 25 26 27 30 31 32 33 35 36 39 40 43 44 45 46 47 48 49 50 57 58 59 62 63 64 66 68 69 70 71 72 73 77	17 19	65

Note. Baseline results are highlighted. Undecided = items eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

Table 26

Consistency in Dimensional-Based Item Clusters between Item-Level EFA and MIRT Cluster Analysis

			Item-Level EFA		MIRT Cluster Analysis		Overall Consistency Index (φ)	
Exam	Form	*N*	*J*	*m*	*J*	*m*	Original	Adjusted
English	2011	500	57	3	56	2	0.66	0.69
		1,000	57	3	56	2	0.64	0.82
		3,000	57	3	56	2	0.67	0.70
	2012	3,000	58	3	58	4	0.65	0.62
Spanish	2011	500	74	4	74	4	0.69	0.69
		1,000	74	4	73	4	0.73	0.73
		3,000	74	4	73	4	0.70	0.70
	2012	3,000	74	4	71	2	0.52	0.65
Chemistry	2011	500	81	1	80	4	0.38	—
		1,000	81	1	81	4	0.64	—
		3,000	81	1	81	4	0.51	—
	2012	3,000	81	1	80	3	0.47	—

Note. Dash line indicates that some stopping rules in cluster analysis cannot examine unidimensionality.
J = number of items actually used in the dimensionality assessment.
m = number of clusters suggested by the method. Original φ = based on the original cluster solutions suggested by the two methods.
Adjusted φ = based on the cluster solutions after adjusting the number of dimensions used in MIRT cluster analysis to equal that proposed by item-level EFA.

Table 27

Dimensionality-Based Item Clusters Produced by Item-Level EFA and MIRT Cluster Analysis for English, Baseline (2011 Form, N = 3,000)

				Dimension	
Item	Format	Content Domain	Testlet	Item-Level EFA	MIRT Cluster Analysis
1	MC	I	R1	**1**	1
2	MC	I	R1	**1**	1
3	MC	I	R1	**1**	1
4	MC	I	R1	**1**	1
5	MC	I	R1	**1**	1
6	MC	I	R1	**1**	1
7	MC	I	R1	**2**	1
8	MC	I	R1	2	1
9	MC	I	R1	**1**	1
10	MC	I	R1	2	1
11	MC	I	R1	**1**	1
12	MC	I	R2	**1**	1
13	MC	I	R2	**1**	1
14	MC	I	R2	**1**	1
15	MC	I	R2	2	1
16	MC	I	R2	1	1
17	MC	I	R2	**1**	1
18	MC	I	R2	1	1
19	MC	I	R2	2	1
20	MC	I	R2	2	1
21	MC	I	R2	**2**	1
22	MC	I	R2	1	1
23	MC	I	R2	1	1
24	MC	I	R2	1	1
25	MC	I	R2	1	1
26	MC	I	R3	**2**	1
27	MC	I	R3	2	1
28	MC	I	R3	**1**	1
29	MC	I	R3	**2**	1
30	MC	I	R3	2	—
31	MC	I	R3	**2**	1
32	MC	I	R3	**2**	1
33	MC	I	R3	**2**	1
34	MC	I	R3	**1**	1
35	MC	I	R3	**2**	1

Note. For item-level EFA, the dimension to which an item was substantially related is emphasized in bold.
R = reading testlet; W = writing prompts (no testlet).

Table 27 *(Cont.)*

Item	Format	Content Domain	Testlet	Dimension	
				Item-Level EFA	MIRT Cluster Analysis
36	MC	I	R3	**2**	2
37	MC	I	R3	**2**	2
38	MC	I	R3	2	1
39	MC	I	R3	**2**	2
40	MC	I	R3	**2**	1
41	MC	I	R3	2	2
42	MC	II	R4	**3**	2
43	MC	II	R4	**3**	2
44	MC	II	R4	**3**	2
45	MC	II	R4	**2**	2
46	MC	II	R4	**3**	2
47	MC	II	R4	**3**	2
48	MC	II	R4	**3**	2
49	MC	II	R4	**3**	2
50	MC	II	R4	**3**	2
51	MC	II	R4	**3**	2
52	MC	II	R4	**3**	2
53	MC	II	R4	**3**	2
54	MC	II	R4	**3**	2
55	CR	I	W	**1**	1
56	CR	I	W	**1**	1
57	CR	II	W	**1**	1

Note. For item-level EFA, the dimension to which an item was substantially related is emphasized in bold.
R = reading testlet; W = writing prompts (no testlet).

Table 28

Dimensionality-Based Item Clusters Produced by Item-Level EFA and MIRT Cluster Analysis for Spanish, Baseline (2011 Form, N = 3,000)

				Dimension	
Item	Format	Content Domain	Testlet	Item-Level EFA	MIRT Cluster Analysis
1	MC	I	L1	**1**	1
2	MC	I	L1	**1**	1
3	MC	I	L1	1	4
4	MC	I	L1	**2**	4
5	MC	I	L2	1	4
6	MC	I	L2	**1**	4
7	MC	I	L2	1	4
8	MC	I	L2	**2**	2
9	MC	I	L2	1	1
10	MC	I	L3	**2**	2
11	MC	I	L3	**2**	2
12	MC	I	L3	**2**	2
13	MC	I	L3	**2**	2
14	MC	I	L4	**1**	1
15	MC	I	L4	**1**	1
16	MC	I	L4	2	—
17	MC	I	L4	**2**	2
18	MC	I	L4	2	4
19	MC	I	L5	1	4
20	MC	I	L5	**2**	2
21	MC	I	L5	2	4
22	MC	I	L5	**2**	4
23	MC	I	L5	**2**	2
24	MC	I	L5	**2**	2
25	MC	I	L5	**2**	2
26	MC	I	L5	**1**	1
27	MC	I	L6	1	4
28	MC	I	L6	2	2
29	MC	I	L6	**1**	4
30	MC	I	L6	**1**	4
31	MC	I	L6	2	2
32	MC	I	L6	1	4
33	MC	I	L6	**1**	4
34	MC	I	L6	**1**	4

Note. For item-level EFA, the dimension to which an item was substantially related is emphasized in bold.
L = listening testlet; R = reading testlet; W = writing prompt (no testlet); S = speaking prompt (no testlet).

Table 28 *(Cont.)*

Item	Format	Content Domain	Testlet	Dimension	
				Item-Level EFA	MIRT Cluster Analysis
35	MC	II	R1	1	4
36	MC	II	R1	**2**	4
37	MC	II	R1	**2**	4
38	MC	II	R1	**2**	2
39	MC	II	R1	**2**	2
40	MC	II	R1	**2**	2
41	MC	II	R1	**2**	4
42	MC	II	R1	**1**	1
43	MC	II	R1	**2**	2
44	MC	II	R2	**1**	1
45	MC	II	R2	**1**	1
46	MC	II	R2	**1**	1
47	MC	II	R2	**1**	1
48	MC	II	R2	**1**	1
49	MC	II	R2	**1**	1
50	MC	II	R2	**1**	1
51	MC	II	R2	**1**	1
52	MC	II	R2	**1**	1
53	MC	II	R3	**4**	4
54	MC	II	R3	**4**	4
55	MC	II	R3	**4**	4
56	MC	II	R3	**4**	4
57	MC	II	R3	**4**	4
58	MC	II	R3	**4**	4
59	MC	II	R3	**4**	4
60	MC	II	R3	4	4
61	MC	II	R3	**4**	4
62	MC	II	R4	**3**	3
63	MC	II	R4	**3**	4
64	MC	II	R4	**3**	3
65	MC	II	R4	**3**	4
66	MC	II	R4	**3**	3
67	MC	II	R4	**3**	4
68	MC	II	R4	**3**	3
69	MC	II	R4	**3**	3
70	MC	II	R4	**3**	3

Note. For item-level EFA, the dimension to which an item was substantially related is emphasized in bold.
L = listening testlet; R = reading testlet; W = writing prompt (no testlet); S = speaking prompt (no testlet).

	Method 2		
Method 1	Items Apart (0)	Item Together (1)	Marginals
Items Apart[a] (0)	φ_{00}	φ_{01}	φ_{0*}
Items Together[b] (1)	φ_{10}	φ_{11}	φ_{1*}
Marginals	φ_{*0}	φ_{*1}	1

Figure 1. Proportions of consistent and inconsistent structural identifications.

Note. [a] Two items assigned to different dimensions by a specific method.

[b] Two items assigned to the same dimension by a specific method.

	MIRT Cluster Analysis		
Item-Level EFA	Items Apart (0)	Item Together (1)	Marginals
Items Apart (0)	0.47	0.09	0.56
Items Together (1)	0.27	0.18	0.45
Marginals	0.74	0.27	1

Figure 2. Example proportions of consistent and inconsistent structural identifications.

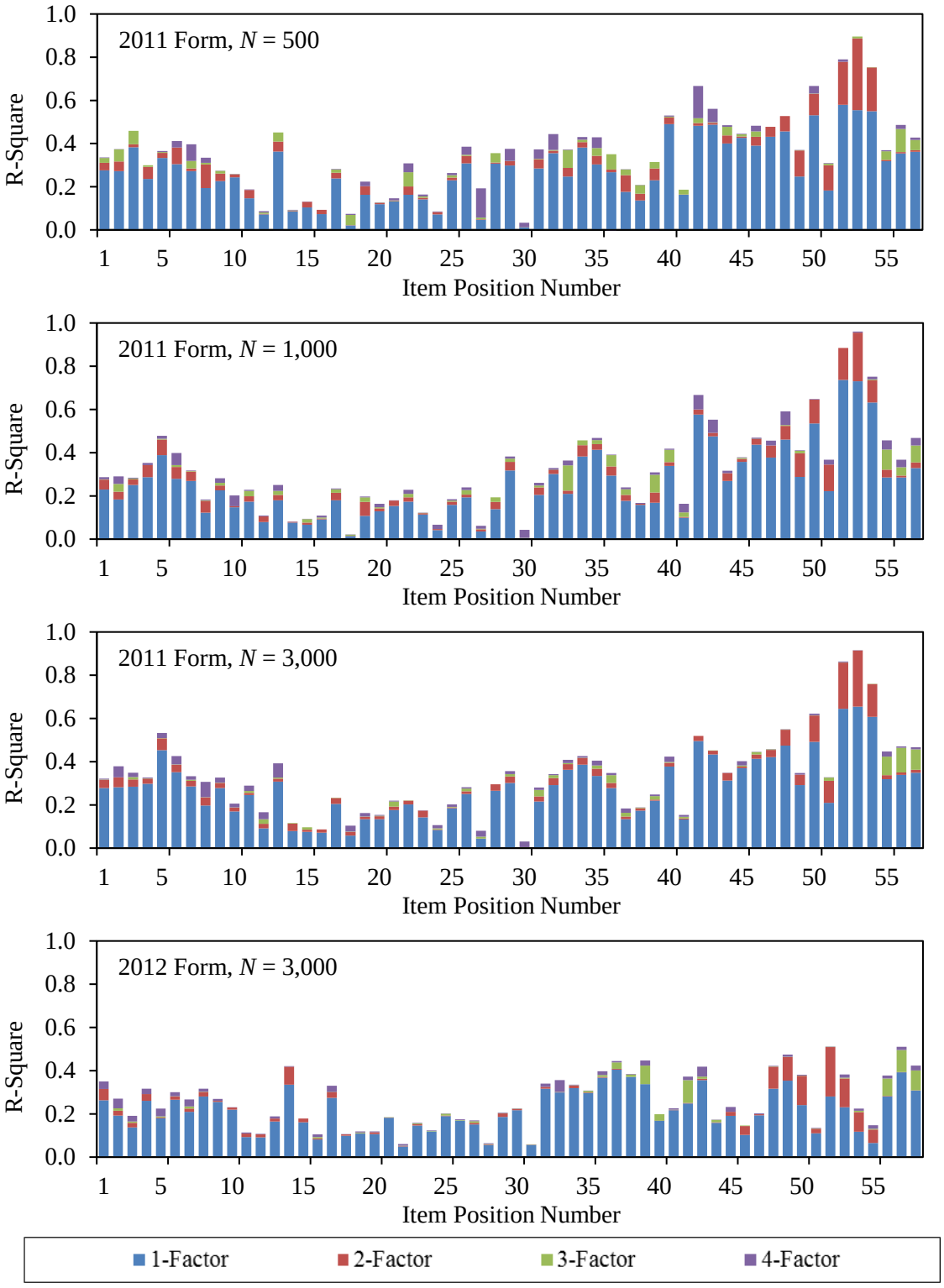


Figure 3. R-squares of 1-factor model and R-square improvements of 2-, 3-, and 4-factor models for the English datasets.

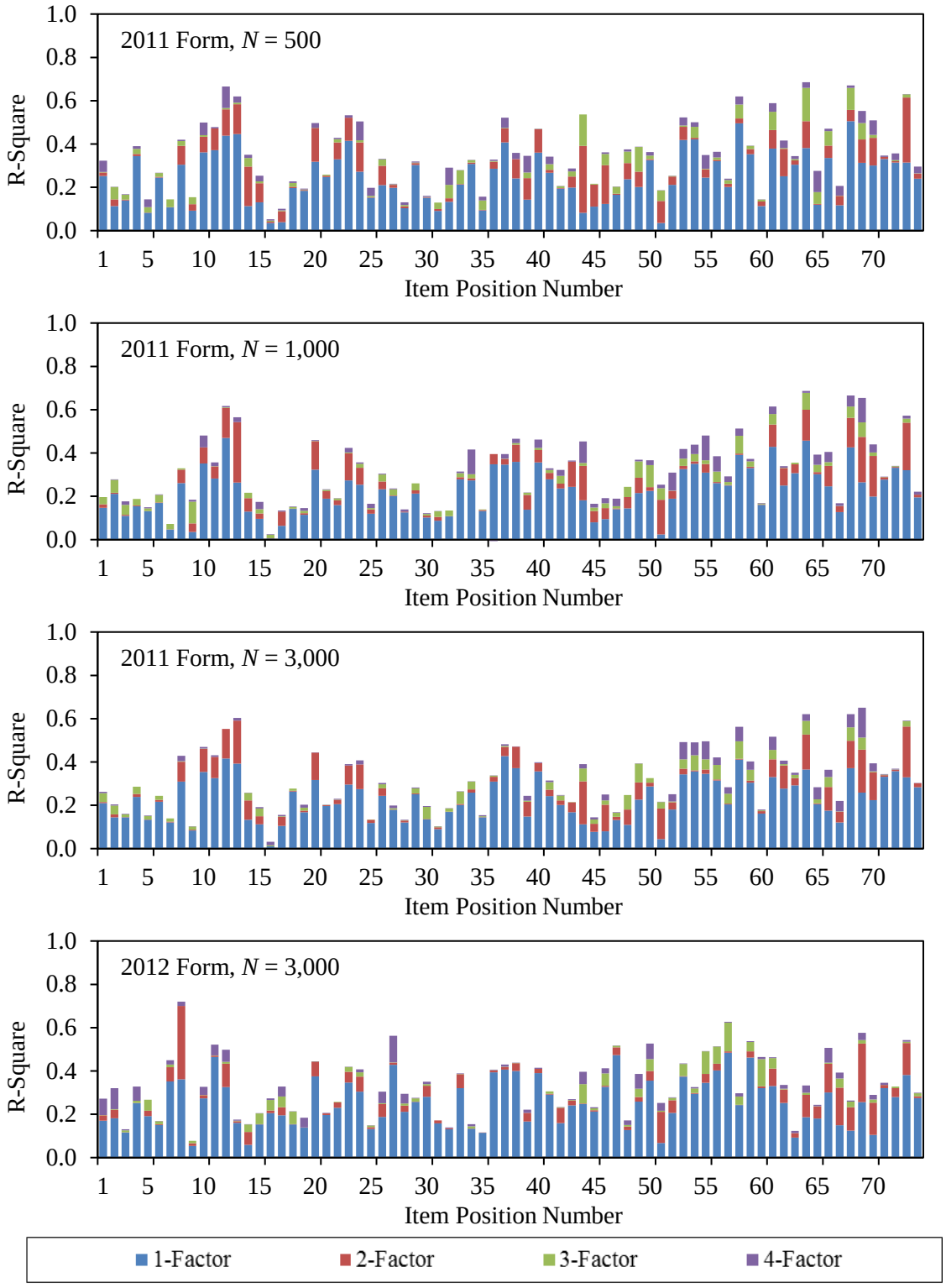


Figure 4. R-squares of 1-factor model and R-square improvements of 2-, 3-, and 4-factor models for the Spanish datasets.

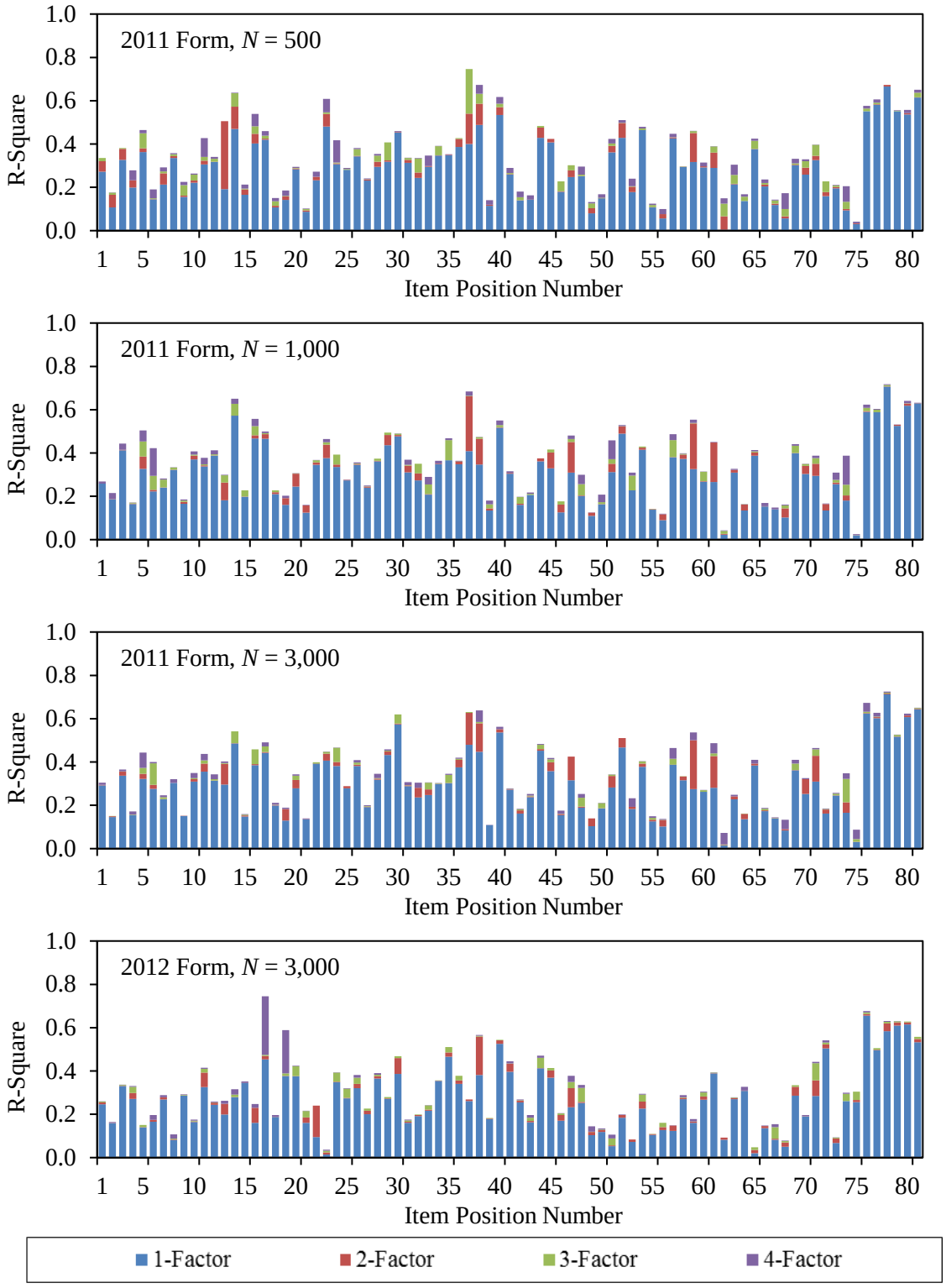


Figure 5. R-squares of 1-factor model and R-square improvements of 2-, 3-, and 4-factor models for the Chemistry datasets.

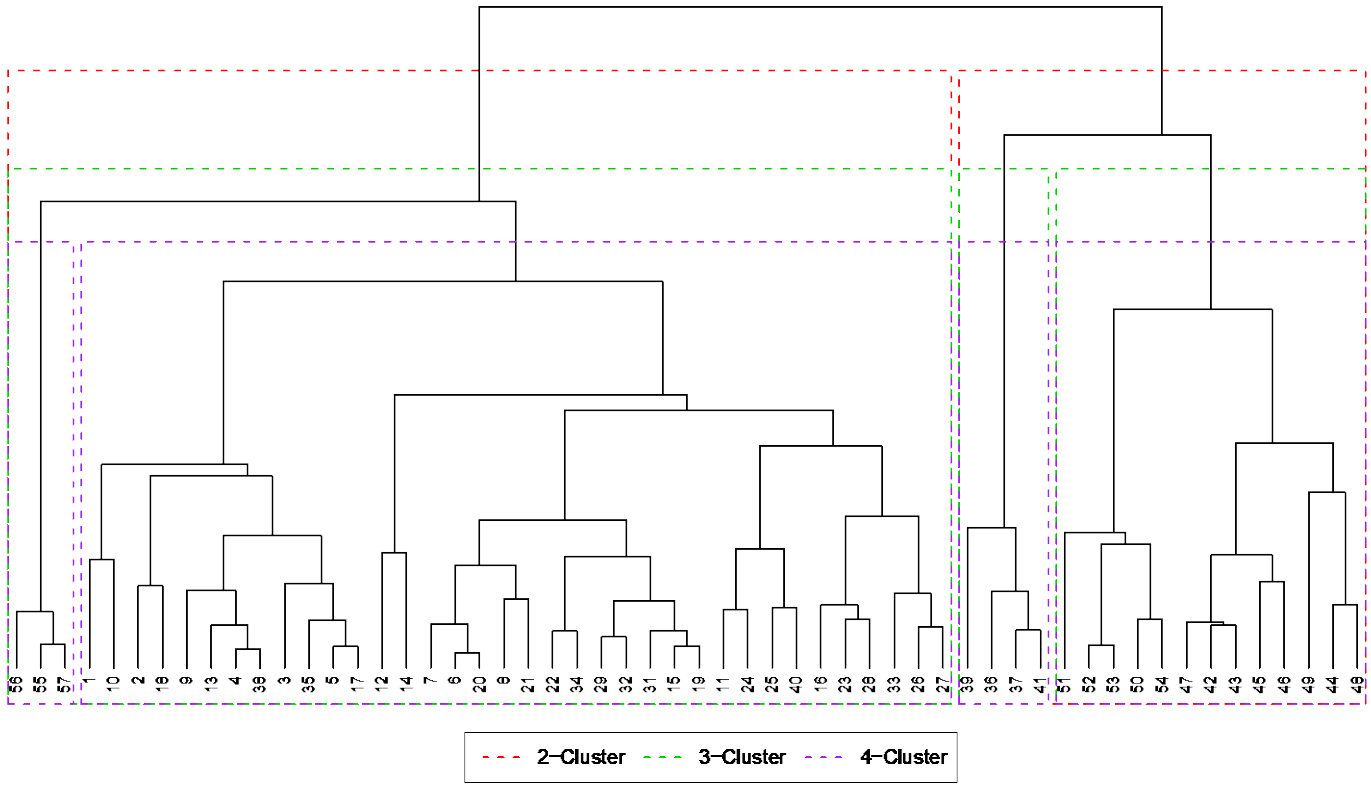


Figure 6. Cluster dendrogram for the English exam, baseline (2011 Form, *N* = 3,000).

Note. Item 30 was eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

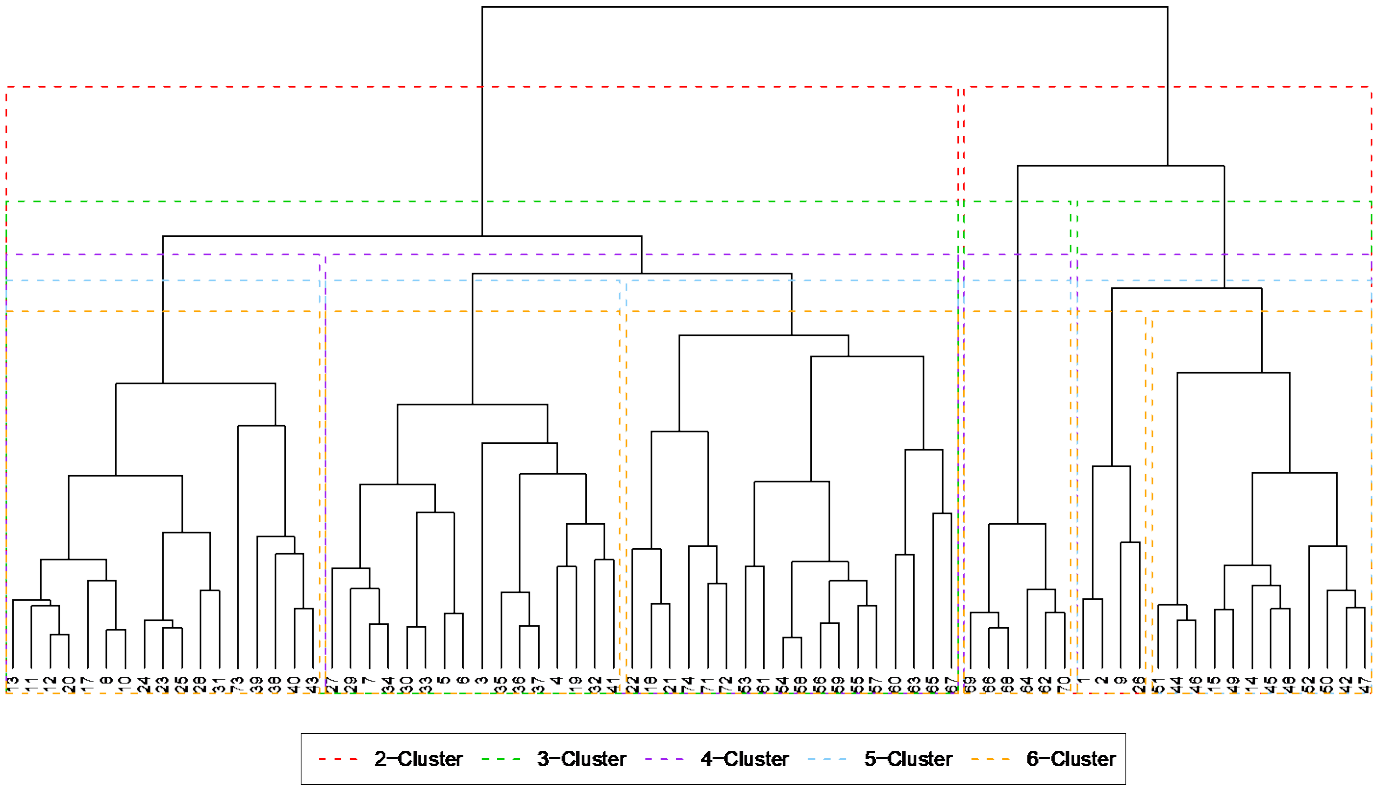


Figure 7. Cluster dendrogram for the Spanish exam, baseline (2011 Form, N = 3,000).

Note. Item 16 was eliminated from MIRT calibration and cluster analysis due to irregular item difficulty and/or discrimination.

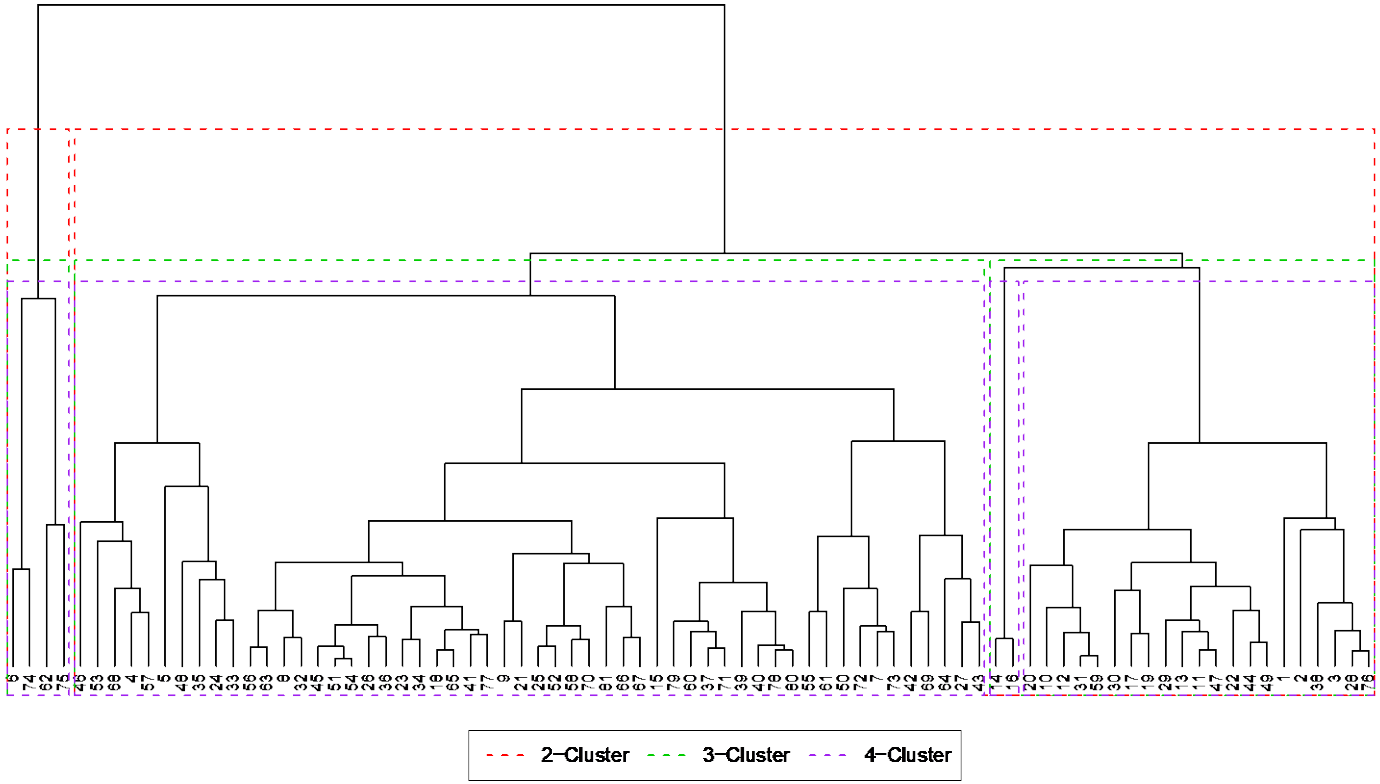


Figure 8. Cluster dendrogram for the Chemistry exam, baseline (2011 Form, *N* = 3,000).

Chapter 8. Linking Methods for the Full-Information Bifactor Model Under the Common-Item Nonequivalent Groups Design

Kyung Yong Kim[1] and Won-Chan Lee[2]

[1]University of North Carolina at Greensboro, Greensboro, NC

[2]University of Iowa, Iowa City, IA

Abstract

Three decisions must be made when using item response theory (IRT) calibration programs to estimate item parameters for multidimensional IRT (MIRT) models: (a) the location of the origin, (b) the unit of measurement along each coordinate axis, and (c) the orientation of the coordinate axes. To handle these three indeterminacies, calibration programs typically set the probability density of the latent variables to a standard multivariate normal distribution. However, by doing so, item parameter estimates obtained from two independent calibration runs with nonequivalent groups result in two different coordinate systems. To achieve a common scale, various linking methods have been introduced and studied for unidimensional IRT models and the full MIRT model. However, little research has been conducted on the linking methods for the full-information bifactor model, although it has been increasingly recognized as a model for fitting educational and psychological data. Thus, the purpose of this study was to provide detailed descriptions of the separate and concurrent calibration methods for the bifactor model and to compare the two linking methods using a simulation study. In general, the concurrent calibration method performed better than the separate calibration method, demonstrating better recovery of the item parameters, test characteristic surface, and expected observed-score distribution.

Linking Methods for the Full-Information Bifactor Model Under the Common-Item Nonequivalent Groups Design

In item response theory (IRT), two major difficulties arise in item parameter estimation. First, the use of the marginal maximum likelihood estimation (MMLE) procedure, implemented via EM algorithm (Bock & Aitkin, 1981; Harwell, Baker, & Zwarts, 1988; Woodruff & Hanson, 1997) or the marginalized Bayesian approach (Harwell & Baker, 1991; Mislevy, 1986), entails the estimation of item parameters by marginalizing (integrating) out the person parameters. This type of estimation requires the person parameters to be described by a probability distribution; however, correctly specifying the distribution of person parameters is often difficult, if not impossible, because they are latent variables. The other difficulty related to item parameter estimation in IRT is the lack of a standard coordinate system. More specifically, in both unidimensional IRT (UIRT) and multidimensional IRT (MIRT), the location of the origin and the unit of measurement along each coordinate axis are arbitrary, and the orientation of the coordinate axes is arbitrary as well in MIRT. To resolve these indeterminacies, IRT calibration programs typically set the distribution of latent variables to a univariate or multivariate standard normal distribution, depending on the number of user-specified dimensions in the IRT model used for calibration.

For the common-item nonequivalent groups (CINEG) design, two test forms with a set of common items are administered to samples from two different populations. This data collection design is often used for test equating when testing programs cannot administer more than one test from on a given test date because of test security concerns. The CINEG design is also widely used in vertical scaling to place tests that differ in difficulty but are intended to measure similar constructs (e.g., math tests across different grade levels) on the same scale. In the context of IRT, the two populations are different in the sense that their distributions of latent variables are different. Nonetheless, when calibrating items for the two nonequivalent groups using two independent computer runs, the distribution of latent variables for each group is often set to a univariate or multivariate standard normal distribution to handle the indeterminacies. Consequently, the two sets of item parameter estimates obtained from two independent calibration runs on nonequivalent groups will be on two different coordinate systems. To handle this issue and place all the item parameter estimates on a common coordinate system, a process

called *linking* is required. Note that the term *linking* in this paper is used to describe explicitly the process of placing IRT parameters on a common coordinate system.

For UIRT models, various linking procedures have been introduced and used in applications such as vertical scaling, differential item functioning (DIF), computerized adaptive testing, and test equating. Among these linking procedures, the relative performance of the separate and concurrent calibration methods has been compared in numerous studies (Hanson & Béguin, 2002; Kim & Cohen, 1998; Kim & Kolen, 2006; Lee & Ban, 2010; Petersen, Cook, & Stocking, 1983). In the context of multidimensional IRT (MIRT), several studies have extended the UIRT separate calibration procedures to the full MIRT model to investigate the recovery of item parameters (Davey, Oshima, & Lee, 1996; Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2000; Yao, 2011). Through simulation, studies consistently found that the proposed methods recover the item parameters reasonably well. In addition, Simon (2008) compared some of the proposed separate calibration methods to the concurrent calibration method and found that the concurrent calibration method generally performs better than the separate calibration methods.

The compensatory MIRT model, which was used in all of the aforementioned MIRT linking studies, is flexible in the sense that each item can freely load on any dimensions in the model. Despite this flexibility, the computational burden for the compensatory MIRT model increases exponentially with increasing dimensions, which substantially lowers the applicability of this model. An alternative MIRT model that is not only computationally tractable but also flexible enough to represent structures that are commonly found in educational and psychological measurement is the full-information bifactor model (Cai, Yang, & Hansen, 2011; Gibbons et al., 2007; Gibbons & Hedeker, 1992). In educational measurement, the bifactor model has been applied to a variety of areas, including calibration of testlet-based tests (DeMars, 2006), vertical scaling (Li & Lissitz, 2012), DIF (Jeon, Rijmen, & Rabe-Hesketh, 2013), multiple-group analysis (Cai et al., 2011), and test equating (Lee et al., 2015; Lee & Lee, 2016). However, although many of these procedures involve linking, little work has been done to compare the separate and concurrent calibration methods in the context of the bifactor model. Thus, building upon earlier work on linking, the objective of this paper is to (a) provide detailed descriptions of the separate and concurrent calibration procedures for the bifactor model; and (b) compare the relative performance of the two linking methods using a simulation study.

Bifactor Model

The full-information bifactor model gained its popularity due to Gibbons and Hedeker's (1992) derivation of a dimension reduction MMLE-EM technique for binary response data in the form of a normal ogive model. Later, Gibbons et al. (2007) derived a dimension reduction estimation procedure for the normal ogive graded response model when the data had a bifactor structure. Recently, Cai et al. (2011) generalized the bifactor model to relate the latent variables and response data using multidimensional extensions of the unidimensional logistic models. The bifactor model used in this study is the bifactor extension of the unidimensional three-parameter logistic (3PL) model presented in Cai et al. (2011).

In the bifactor model, there are two different sets of latent variables: (a) the general factor and (b) the specific factors. The model requires that all items load on the general factor and on at most one specific factor. Furthermore, it is commonly assumed that all factors are independent to achieve dimension reduction during computations of the marginal maximum likelihood. The following is an example of a bifactor pattern for four items assuming two specific dimensions (the last two columns):

$$\begin{bmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & 0 & a_{32} \\ a_{40} & 0 & a_{42} \end{bmatrix}, \tag{1}$$

where the a's denote nonzero item slopes (i.e., item discrimination parameters). In Equation 1, all four items load on the general factor (the first column), the first two items load on the first specific factor (the second column), and the last two items load on the second specific factor (the third column).

The bifactor extension of the unidimensional 3PL model is expressed as

$$\mathrm{P}\left(\boldsymbol{\theta}_i, a_{j0}, a_{js}, c_j, d_j\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7\left(a_{j0}\theta_{i0} + a_{js}\theta_{is} + d_j\right)\right]}, \tag{2}$$

where the subscripts i and j denote person and item, respectively; θ_{i0} is the general factor; θ_{is} is the specific factor on which item j loads; $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{is})$; a_{j0} and a_{js} are the item discrimination parameters for the general and specific factors, respectively; c_j is the pseudo-guessing parameter; and d_j is the intercept parameter that is related to the multidimensional difficulty parameter $\left(\mathrm{MDIFF}_j\right)$. Adopting Reckase's (2009) formula, the relationship between d_j and MDIFF_j is

$$\text{MDIFF}_j = -\frac{d_j}{\sqrt{a_{j0}^2 + a_{js}^2}}\,. \tag{3}$$

When $a_{js} = 0$, Equation 3 becomes the unidimensional difficulty parameter, and the bifactor model reduces to the unidimensional three-parameter logistic (3PL) model.

Separate and Concurrent Calibration for the Bifactor Model

This section provides detailed descriptions of the separate and concurrent calibration procedures for the bifactor model. It is assumed that the purpose of linking is to place the item parameters of the new form on the coordinate system of the base form. Note that "Form X" will be used interchangeably with "new form", and "Form Y" will be used interchangeably with "base form". In addition, the groups that take Forms X and Y will be referred to as the new and base groups, respectively.

Under the CINEG design, three decisions must be made when using IRT calibration programs to place item parameter estimates obtained from separate calibrations on the same coordinate system for MIRT models: (a) the location of the origin, (b) the unit of measurement along each coordinate axis, and (c) the orientation of the coordinate axes, which is often referred to in the literature as rotational indeterminacy. Note that rotational indeterminacy for the bifactor model is fixed under the standard bifactor analysis assumption of complete independence of all dimensions. However, as noted by Rijmen (2009), dimension reduction in likelihood computations can also be achieved by relaxing the complete independence assumption to conditional independence of the specific factors given the general factor. In this case, the rotational indeterminacy can be fixed by constraining the correlation between the general factor and each specific factor to zero. The separate and concurrent calibration procedures presented in this section are based on the standard bifactor analysis assumption of complete independence.

Separate Calibration

Extending the transformation equations for the 3PL model (Kolen & Brennan, 2014, p. 178) to the bifactor model, the item response probabilities remain unchanged with the following linear transformations:

$$\boldsymbol{\theta}_{Y_i} = \boldsymbol{C}^{-1}\left(\boldsymbol{\theta}_{X_i} - \boldsymbol{\beta}\right), \tag{4}$$

$$\boldsymbol{a}_{Y_j}^T = \boldsymbol{a}_{X_j}^T \boldsymbol{C}, \tag{5}$$

$$d_{Y_j} = d_{X_j} + \boldsymbol{a}_{X_j}^T \boldsymbol{\beta}, \tag{6}$$

and

$$c_{Y_j} = c_{X_j}, \tag{7}$$

where $\boldsymbol{C}$ is a diagonal matrix of size $(S+1) \times (S+1)$ with scaling constants at the diagonals (S denotes the total number of specific factors in the model); $\boldsymbol{\beta}$ is the translation vector of length $S+1$; $\boldsymbol{\theta}_{Y_i}$ is the person parameter vector for person i on the base scale; $\boldsymbol{a}_{Y_j}$, c_{Y_j}, and d_{Y_j} are the slope parameter vector, pseudo-guessing parameter, and intercept parameter for item j on the base scale, respectively; and $\boldsymbol{\theta}_{X_i}$, $\boldsymbol{a}_{X_j}$, c_{X_j}, and d_X are the corresponding parameters on the scale of the new form. Probability invariance can be shown mathematically by comparing the exponent of the bifactor model before transformation to that after transformation:

$$\begin{aligned}\boldsymbol{a}_{Y_j}^{\mathrm{T}}\boldsymbol{\theta}_{Y_i} + d_{Y_j} &= \left(\boldsymbol{a}_{X_j}^{\mathrm{T}}\boldsymbol{U}\right)\left(\boldsymbol{U}^{-1}\left(\boldsymbol{\theta}_{X_i} - \boldsymbol{\beta}\right)\right) + \left(d_{X_j} + \boldsymbol{a}_{X_j}^{\mathrm{T}}\boldsymbol{\beta}\right) \\ &= \boldsymbol{a}_{X_j}^{\mathrm{T}}\left(\boldsymbol{\theta}_{X_i} - \boldsymbol{\beta}\right) + \left(d_{X_j} + \boldsymbol{a}_{X_j}^{\mathrm{T}}\boldsymbol{\beta}\right) = \boldsymbol{a}_{X_j\boldsymbol{\theta}X_i}^{\mathrm{T}} + d_{X_j}.\end{aligned} \tag{8}$$

Because the scaling matrix $\boldsymbol{C}$ is diagonal and only appears in Equation 5 in the context of the CINEG design, one way to estimate the scaling parameter for each dimension is to divide the mean of the slope parameter estimates for the base form by that of the new form as follows:

$$\hat{c}_h = \frac{\mathbf{E}\left(\hat{a}_{Y_h}\right)}{\mathbf{E}\left(\hat{a}_{X_h}\right)}, \qquad h = 0, 1, \ldots, S, \tag{9}$$

where **E** is the expectation operator. Note that only the slope parameter estimates for the common items are used for computing each $\hat{c}_h$. The translation vector can be estimated by solving Equation 6 for $\boldsymbol{\beta}$; that is,

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{A}}_{X_c}^{T}\widehat{\boldsymbol{A}}_{X_c}\right)^{-1}\widehat{\boldsymbol{A}}_{\mathrm{X_c}}^{\mathrm{T}}\left(\widehat{\boldsymbol{d}}_{Y_c} - \widehat{\boldsymbol{d}}_{X_c}\right), \tag{10}$$

where $\boldsymbol{A}_{X_c}$ is a matrix of the slope parameters for the common items, and $\boldsymbol{d}_{Y_c}$ and $\boldsymbol{d}_{X_c}$ are vectors of the intercept parameters for the common items in the base and new forms, respectively.

Concurrent Calibration

As described earlier, when estimating item parameters for multiple test forms that are administered to nonequivalent groups, the parameter estimates that result from separate runs of an IRT calibration program are not on the same coordinate system because, for each computer run, the distribution of latent variables is arbitrarily set to a multivariate standard normal distribution. In concurrent calibration, this scale problem is handled by estimating the item parameters and the distributions of the latent variables for the base and new groups

simultaneously. For concurrent calibration, the contribution of examinee i in group g to the marginal likelihood is

$$L(\boldsymbol{\Delta} \mid \boldsymbol{u}_{i(g)}) = \int_R \int_{R^S} f(\boldsymbol{u}_{i(g)} \mid \boldsymbol{\theta}_g, \boldsymbol{\Delta}) h(\boldsymbol{\theta}_g \mid \boldsymbol{\Delta})\, d\boldsymbol{\theta}_g\,, \tag{11}$$

where $\boldsymbol{\Delta}$ is the set of all the item and structural parameters (i.e., parameters related to the distribution of latent variables) in the model; $\boldsymbol{u}_{i(g)} = \left(u_{i(g)1}, \ldots, u_{i(g)J_g}\right)$ is the item responses for examinee i in group g for all J_g items; $\boldsymbol{\theta}_g$ is a vector of the general and specific factors for group g; and $h(\boldsymbol{\theta}_g|\boldsymbol{\Delta})$ is the distribution of latent variables for group g. Also,

$$f(\boldsymbol{u}_{i(g)} \mid \boldsymbol{\theta}_g, \boldsymbol{\Delta}) = \prod_{s=1}^{S} \prod_{j \in I_g \cap I_s} f(u_{i(g)j} \mid \theta_{g0}, \theta_{gs}, \boldsymbol{\Delta})\,, \tag{12}$$

where θ_{g0} is the general factor for group g; θ_{gs} is the specific factor for group g on which item j loads; $f(u_{i(g)j} \mid \theta_{g0}, \theta_{gs}, \boldsymbol{\Delta})$ is the item response function for the bifactor model; I_g denotes the set of items that is given to the examinees in group g; I_s denotes the set of items that load on specific factor θ_{gs}; and $\cap$ is the intersection operator. As implied by the term $j \in I_g \cap I_s$, Equation 11 is computed using only the items that are taken by group g. Because the latent variables in the bifactor model are orthogonal by the standard bifactor analysis assumption and each item loads on at most one specific factor, Equation 11 can be rewritten as

$$L(\boldsymbol{\Delta} \mid \boldsymbol{u}_{i(g)}) = \int_R \prod_{s=1}^{S} \left[\int_R \prod_{j \in I_g \cap I_s} f(u_{i(g)j} \mid \theta_{g0}, \theta_{gs}, \boldsymbol{\Delta}) h(\theta_{gs} \mid \boldsymbol{\Delta})\, d\theta_{gs} \right] g(\theta_{g0} \mid \boldsymbol{\Delta}) d\theta_{g0}\,. \tag{13}$$

Invoking the local independence assumption, the (overall) marginal log-likelihood function becomes

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} \log L(\boldsymbol{\Delta} \mid \boldsymbol{u}_{i(g)})\,, \tag{14}$$

where G is the number of groups, and N_g is the number of examinees in group g.

Extending the EM algorithm used for UIRT models (Bock & Aitkin, 1981; Harwell, Baker, & Zwarts, 1988; Woodruff & Hanson, 1997) to the bifactor model, the E-step of the EM algorithm involves creating three types of “artificial data” for each group using item responses

and provisional item parameter estimates obtained from the previous EM cycle. The three types of "artificial data" are (a) the expected number of examinees in group g at the Q quadrature points for the general dimension, $\{X_{q_0}\}_{q_0=1}^{Q}$; (b) the expected number of examinees in group g at the Q quadrature points for each specific dimension, $\{X_s\}_{q_s=1}^{Q}$; and (c) the expected number of examinees in group g that responds to score category k (either 0 or 1 for a dichotomous item) of item j at each combination of (X_{q_0}, X_{q_s}). These three quantities are denoted hereinafter by $r_g(X_{q_0})$, $r_{gs}(X_{q_s})$, and $r_{gjk}(X_{q_0}, X_{q_s})$, and can be computed using a method similar to that described in Cai (2010) for the two-tier model.

In the M-step, the parameter estimates for each item are updated by finding the values that maximize

$$\phi_j(\boldsymbol{\Delta}) = \sum_{g=1}^{G}\left[I_{j\in I_g}\sum_{q_0=1}^{Q}\sum_{q_s=1}^{Q}\sum_{k=1}^{K} r_{gjk}(X_{q_0}, X_{q_s}) \log f\left(u_{i(g)j} \mid \theta_{g0}, \theta_{gs}, \boldsymbol{\Delta}\right)\right], \qquad (15)$$

where K is the number of score categories, and $I_{j\in g}$ is an indicator variable such that $I_{j\in g} = 1$ if $j \in g$ and otherwise $I_{j\in g} = 0$. Because the first summation is over groups, item responses for all the groups that take item j (i.e., $I_{j\in g} = 1$) are used to maximize Equation 14. This is the reason that the concurrent calibration method provides a single set of parameter estimates for the common items.

In addition to the item parameters, the probability distributions of the latent variables for all G groups are updated in the M-step as well. When assuming that the general and specific factors for each of the G groups jointly follow a multivariate normal distribution, estimating the probability distribution of each group is equivalent to estimating the mean vector and covariance matrix of the distribution. Adopting the standard bifactor analysis assumption that the general and specific factors are orthogonal, each element of the mean vector and covariance matrix can be estimated separately under the assumption that each latent variable is normally distributed. Specifically, the first two moments of the normal distribution for the general and specific factors can be estimated by finding the structural parameters (i.e., the mean and standard deviation) that maximize

$$\psi_g(\boldsymbol{\Delta}) = \sum_{q_0=1}^{Q} r_g\left(X_{q_0}\right) \log f_g\left(X_{q_0} \mid \boldsymbol{\Delta}\right), \tag{16}$$

and

$$\psi_{gs}(\boldsymbol{\Delta}) = \sum_{q_s=1}^{Q} r_{gs}\left(X_{q_s}\right) \log f_{gs}\left(X_{q_s} \mid \boldsymbol{\Delta}\right), \tag{17}$$

respectively, where $f_g\left(X_{q_0} \mid \boldsymbol{\Delta}\right)$ and $f_{gs}\left(X_{q_s} \middle| \boldsymbol{\Delta}\right)$ are the probability density functions of a univariate normal distribution. Once the latent variable distributions for all G groups are updated, the distribution for the reference group is linearly transformed to have a specific mean and standard deviation (e.g., a mean vector of **0** and a covariance matrix of **I**). The purpose of doing so is to fix the location of the origin and unit of measurement along each coordinate axis of the coordinate system. Then, the same linear transformation is applied to the distributions for the other groups to ensure that the relative location of all the distributions remain unchanged. Furthermore, the item parameter estimates should also be linearly transformed in such a way that the probability of correct response remains unchanged. These revised latent variable distributions and item parameter estimates are used then in the following EM cycle.

Instead of assuming that the latent variables are normally distributed, the distributions for all G groups can be estimated empirically by adopting the empirical histogram method for UIRT models (Mislevy, 1984; Woodruff & Hanson, 1997). This approach is relatively easy to implement because empirical histograms are by-products of the "artificial data" computed in the E-step of the EM algorithm; that is, the empirical histograms for the general and specific factors for each group can be obtained by simply dividing $r_g\left(X_{q_0}\right)$ and $r_{gs}\left(X_{q_s}\right), s = 1, \ldots, S$, by the total number of examinees in that group.

Method

The relative performance of the separate and concurrent calibration methods was compared based on a simulation study. To obtain more stable item parameter estimates, all analyses were conducted using a bifactor model with no pseudo-guessing parameter. An *R* (R Core Team, 2017) program that was written by the author of this study was used to implement the separate and concurrent calibration procedures.

Study Factors

Two study factors were included in the simulation study: (a) two levels of sample size (*N* = 1,000 and 2,000), and (b) two levels of the proportion of common items (CI = 20% and 40%). The two factors were completely crossed, resulting in four study conditions. The data generation procedure is described in the following section; for each condition, 100 response datasets were generated by repeating the data generation procedure 100 times.

Simulation Procedures

Two dichotomously scored 45-item test forms with three categories, each category having 15 items, were generated for the simulation study. To create the two test forms, item parameters for the bifactor model were sampled from probability distributions that are commonly used in IRT studies, and the parameters of the probability distributions were chosen such that the first two moments of the generated item parameters were close to those used in previous studies (Cai, 2011; DeMars, 2006; Li & Lissitz, 2012). More specifically, the general slope parameters (a_0) were sampled from a normal distribution with a mean of .9 and a standard deviation of .2; three sets of specific slope parameters (a_s), one for each category, were sampled from a uniform distribution between .5 and .7; and the multidimensional difficulty parameters (MDIFF) were sampled from a standard normal distribution for the base form and for the new form using a normal distribution with a mean of 0.5 and a standard deviation of 1. The reasoning behind sampling multidimensional difficulty parameters from normal distributions with different mean values for the two groups was to introduce form differences. Using Reckase's (2009, p. 90) formula, the multidimensional difficulty parameters were converted to intercept parameters (d). These item parameters will be hereafter referred to as the generating item parameters.

Using the generating item parameters, item responses for the base form were generated by sampling the general and each of the specific factors from a $\mathrm{N}(0, 1)$ distribution. For the new form, item responses were generated by sampling the general factor from a $\mathrm{N}(1, 1)$ and the three specific factors from N(0, 1), N(0.5, 1), and N(1, 1) distributions, respectively. The reason for using three different distributions for the specific factors was to examine the recovery of item parameters under a variety of plausible conditions.

To introduce two different proportions of common items, parameters for 15 items in each category were sampled in such a way that six items had statistical characteristics that were comparable to all 15 items. These six items were considered as common items for the 40%

common-item condition. For the 20% common-item condition, the original six common items were split into two sets such that the statistical characteristics for each set of three items were as similar as possible. One of the two sets of three items was considered as common items, and the other set was treated as unique items.

Evaluation Criteria

Four evaluation criteria were used to compare the three linking methods. First, the recovery of the transformation parameters (i.e., the scaling and translation parameters) for the separate calibration method and the first four moments of the distribution of latent variables for the concurrent calibration method were examined. The first four moments were considered for the concurrent calibration method because the empirical histogram method was used to estimate the distribution of the latent variables rather than assuming normality and estimating only the first two moments. For separate calibration, the translation parameters for the four dimensions were -1, 0, -0.5, and -1, and the scaling parameters for all four dimensions were 1. For concurrent calibration, the mean vector of the latent variable distribution was (-1, 0, -0.5, -1), the covariance matrix was the identity matrix, the skewness for all four dimensions was 0, and the kurtosis for all four dimensions was 3.

After conducting linking, the item parameter estimates for the new form should be close to the generating item parameters. The extent to which this holds was assessed by three criteria. The first criterion compared the linked item parameter estimates to the generating item parameters in terms of bias, standard error (SE), and root mean square error (RMSE). Only the new form items were used for the comparison because the focus of this paper was to place the item parameter estimates for the new form on the coordinate system of the base form. Bias, which measures the systematic error of an estimator, was defined by

$$\text{Bias}_j = \frac{1}{R}\sum_{r=1}^{R} \hat{\nu}_{jr} - \nu_j, \tag{18}$$

where $R\ (= 100)$ is the number of replications, $\hat{\nu}_{jr}$ denotes an estimate for item j at replication r, and ν_j denotes a parameter for the same item. A small value of bias indicates that the mean of an estimator is not systematically different from the parameter value. SE quantifies sampling error and was expressed as

$$\mathrm{SE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{v}_{jr} - \bar{\hat{v}}_j\right)^2}, \tag{19}$$

where $\bar{\hat{v}}_j = \frac{1}{R}\sum_{r=1}^{R}\hat{v}_{jr}$. RMSE takes into account both systematic and random (sampling) error and was defined as the square root of the sum of the squared bias and squared SE:

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{v}_{jr} - v_j\right)^2}. \tag{20}$$

Under each condition, values of bias, SE, and RMSE were first computed for each item, and then the averages were taken over all the items to summarize the results at the test level. For bias, the absolute values were averaged to prevent cancellation of positive and negative values across different items.

The second criterion compared the test characteristic surface (TCS) obtained with the linked item parameter estimates to that obtained with the generating item parameters. As with the first criterion, only the new form TCS was used for comparison. The bias, SE, and RMSE at a given $\boldsymbol{\theta}$ combination for the TCS criterion were defined similarly to Equations 18, 19, and 20, respectively, but substituting $\hat{\tau}_r(\boldsymbol{\theta}) = \sum_{j=1}^{J}\mathrm{P}\left(\boldsymbol{\theta}_i, \hat{a}_{jr0}, \hat{a}_{jrs}, \hat{d}_{jr}\right)$ for $\hat{v}_{jr}$ and $\tau(\boldsymbol{\theta}) = \sum_{j=1}^{J}\mathrm{P}\left(\boldsymbol{\theta}_i, a_{j0}, a_{js}, d_j\right)$ for v_j. The three statistics were averaged over all possible combinations of $\boldsymbol{\theta}$ using two weight functions. The first weight function was $w(\boldsymbol{\theta}) = 1$, and the second weight function was the probability density function for a multivariate standard normal distribution. For each dimension, 21 quadrature points between -4 and 4 were used for the computation.

The last criterion was the expected observed-score distribution (EOD) criterion, which compared the model-based marginal observed-score distribution obtained with the linked item parameter estimates to that obtained with the generating item parameters. For each number-correct score x, the bias, SE, and RMSE for the EOD criterion were computed in a similar way to the TCS criterion using $f(x|\boldsymbol{\theta})$ and $\hat{f}_r(x|\boldsymbol{\theta})$, which denote, respectively, the expected probabilities for number-correct score x at a given $\boldsymbol{\theta}$ combination computed using the generating item parameters and linked item parameter estimates. Note that, unlike the TCS criterion, the EOD criterion is evaluated at each number-correct score x. The Lord-Wingersky recursion

formula (Lord & Wingersky, 1984) was used to compute the expected distributions, and the same two weight functions used for the TCS criterion were also used for the EOD criterion.

Results

For some datasets, calibration runs for the separate and concurrent calibration methods did not converge within 500 EM cycles. These problematic replications were discarded and replaced with new datasets until all 100 datasets converged successfully.

Recovery of Transformation Parameters and Latent Variables Distributions

The top part of Table 1 provides the results for the recovery of the transformation parameters for the separate calibration method. As expected, the transformation parameters and the first four moments were more accurately estimated as either the sample size or the proportion of common items increased. For the separate calibration method, the scaling and translation parameters for all four dimensions were recovered reasonably well, with the general dimension showing estimates that were closer to the population parameters. More accurate estimates were observed for the general dimension due to the number of common items used to estimate the transformation parameters in Equations 9 and 10—nine versus three for the CI = 20% condition, and eighteen versus six for the CI = 40% condition. For all four dimensions, the scaling parameter was recovered more accurately than the translation parameter, showing smaller bias and standard error.

Results for the recovery of the first four moments of the distribution of latent variables for the concurrent calibration method are provided in the bottom part of Table 1. In general, the mean of the general dimension was slightly overestimated, whereas the means of the specific dimensions were systematically underestimated across all four simulation conditions. It is worth noting that the three specific dimensions were underestimated to a similar degree. In contrast to the means, the standard deviations were well recovered for all four dimensions. In terms of skewness and kurtosis, the concurrent calibration method tended to overestimate the population values for all four dimensions. Unlike the separate calibration method, the first four moments for the distribution of the general dimension were not necessarily recovered more accurately than those for the specific dimensions. The reason that the first four moments for the distributions of the general and specific dimensions were recovered equally well was because the common and unique items were both used to estimate the distribution of latent variables for the concurrent calibration method.

Item Parameter Recovery Criterion

Values of average bias (ABIAS), average standard error (ASE), and average root mean square error (ARMSE) for the item parameter recovery criterion are provided in Table 2. Three common findings were observed for the two linking methods. First, the values of ARMSE decreased with increasing sample size. This result was expected because larger sample sizes generally lead to increased precision (i.e., smaller ASE) when estimating population parameters. Second, increasing the proportion of common items tended to lead to lower ARMSE. This result was also expected because different scales are linked through the common items. However, compared to the improvement observed with increasing sample size, the impact of the proportion of common items on the recovery of item parameters on ARMSE was minimal. Third, regarding the magnitude of the generating item parameters, the slope parameters were recovered less accurately for the specific dimensions than for the general dimension.

Comparing the two linking methods, the concurrent calibration method tended to recover the generating item parameters more accurately than the separate calibration method, showing smaller values of ARMSE primarily due to the smaller ASE. The two linking methods returned item parameter estimates that were nearly unbiased, and therefore, the values of ARMSE were almost identical to those of ASE. Among the three specific slope parameters, slopes for θ_1 and θ_2 showed smaller ARMSE than those for θ_3. This was probably because θ_3 had the largest group difference between the base and new groups; θ_1 had no group difference; θ_2 had a group difference of .5; and θ_3 had a group difference of 1.

Results for the recovery of the item parameters across the common and unique items are provided in Table 3. For both linking methods, ARMSE for the common items were smaller than those for the unique items. However, the concurrent calibration method showed larger difference in ARMSE between the common and unique items than the separate calibration method. Furthermore, as ARMSE for the unique items were comparable for the two linking methods, larger difference between the separate and concurrent calibration methods were observed for the common items. This result was somewhat expected because the concurrent calibration method used responses for both the base and new groups to estimate the parameters for the common items, whereas the separate calibration method only uses responses for the new group to estimate the parameters for the common items. It is worth noting these findings are similar to the findings of Hanson and Béguin (2002) in the context of UIRT.

Test Characteristic Surface and Expected Observed-Score Distribution Criteria

Values of bias, SE, and RMSE for the unweighted (uniform weight) and weighted (normal weight) average criteria are presented in Table 4. Because the unweighted and weighted results were similar, the discussion of the results applies to both cases. In contrast to the ICS criterion, the ESD criterion provides bias, SE, and RMSE at each number-correct score. The results presented in Figure 5 are the average over all possible score points. In addition, because the ESD criterion is based on relative frequencies, the values of the three statistics were very small. Therefore, to make meaningful comparison between the three linking methods, the initial values were multiplied by 100.

For the same reason as the item parameter recovery criterion, the values of ARMSE decreased with increasing sample size and the proportion of common items. Comparing the two linking procedures, the concurrent calibration method provided smaller values of ARMSE than the separate calibration method, which was mainly attributable to the smaller SE. However, unlike the item parameter recovery criterion, the concurrent calibration method also provided less biased TCSs and EODs, regardless of the sample size and proportion of common items. Across all four simulation conditions, the separate and concurrent calibration methods showed smaller differences under the normal weight function. The RMSE for the ESD criterion at each score point (i.e., conditional RMSE) is plotted in Figure 6. The general pattern of RMSE was similar for the two linking methods showing an inverted U-shape. Consistent with the overall results, the concurrent calibration method provided the smallest RMSE across a wide range of score points.

Discussion

When using IRT with the CINEG design, IRT linking is a necessary practice in test equating, differential item functioning, vertical scaling, and computerized adaptive testing, among other applications, to place all item parameter estimates on a common scale. Although many research studies compared IRT linking methods for unidimensional IRT models and full MIRT models, it has been an infrequent topic in the literature for the bifactor model. Thus, in this paper, the relative performance of the separate and concurrent calibration methods was compared using a simulation study. Furthermore, descriptions of the separate and concurrent calibration methods were provided to help measurement practitioners understand the specific details of the two linking methods.

Overall, the concurrent calibration method outperformed the separate calibration method, showing better recovery of the item parameters, test characteristic surface, and expected observed-scored distribution. In addition, the concurrent calibration method only requires a single computer run compared to the multiple runs required for the separate calibration method. Despite these advantages, however, the concurrent calibration method requires response data for the base form at the time of calibration, which are often not available. By contrast, the separate calibration method only requires the base form parameter estimates for the common items to conduct linking. Therefore, the separate calibration method would be a more viable option for linking when only the item parameter estimates for the base form are available.

Although the separate calibration method provided larger error than the concurrent calibration method across all the simulation conditions, one distinct advantage of the separate calibration method is that it can be used to examine parameter drift for the common items. Drift can be examined because the separate calibration method provides two separate sets of parameter estimates for the common items, whereas the concurrent calibration method only provides one. For this reason, as suggested by Hanson and Béguin (2002) in the context of UIRT, it would be beneficial in practice to compute the separate calibration estimates for diagnostic purposes, even if the concurrent calibration method is used for operational purposes.

In contrast to the separate calibration method, the concurrent calibration method achieves a common scale for the IRT parameters by estimating the distributions of the latent variables for the base and new forms along with the item parameters. In the context of UIRT, the two most widely used approaches for estimating the probability densities are the normal distribution and empirical histogram methods. Although the use of the normal distribution method appears to be more common, the advantage of the empirical histogram method is that it can be used to handle non-normal latent variables. Another possible approach for estimating the latent variable distribution is to combine the normal distribution and empirical histogram methods. For example, the computer program flexMIRT (Cai, 2017) characterizes the distribution for the general dimension using an empirical histogram, but assumes that the specific dimensions are jointly normally distributed when invoking the EmpHist =Yes option. As noted by Cai et al. (2011), semiparametric approaches suggested in the context of UIRT (e.g., Woods & Thissen, 2006) could also be applied to the bifactor model. It is suspected that different approaches would result in different item parameter estimates. Thus, further research on this topic is warranted.

Besides the separate and concurrent calibration methods, another linking method referred to as the fixed parameter calibration method (Kang & Petersen, 2011; Kim, 2006) in the literature, is often used in practice with UIRT models for linking pretest items to the scale of an item pool. The fixed parameter calibration method fixes the parameter estimates for the common items at their existing values, and estimates the distribution of the latent variables for the new group using the fixed parameter estimates and the responses to the common items obtained from the new group. By doing so, the distribution for the new group and the parameters for the unique items estimated using the estimated distribution is on the desired scale without any further linking steps. As with the separate calibration method, the fixed parameter calibration method only requires the parameter estimates for the base form at the time of calibration. Therefore, it could also be a viable alternative to the concurrent calibration method. Applying the fixed parameter calibration method to the bifactor model could be pursued in future research.

There are several limitations inherent in this study. First, item responses were generated under the assumption that the bifactor model fits the data perfectly. However, this never happens in the real word, which limits the generalizability of the findings of this study. To examine the performance of the separate and concurrent calibration methods for the bifactor model under more realistic conditions, future studies might consider comparing these two linking methods under some degrees of model misfit. Second, in this study, the standard deviations for the base and new groups were both set to one, assuming that the two groups only differed in their means; however, the groups involved in linking might be different in terms of both the mean and standard deviation of the latent variable distribution. In addition, only normal latent variables were considered in this study. A future study might consider generating response data from a skewed distribution to compare the impact of different estimation approaches for the population latent distribution on linking. Last, the effect of the pseudo-guessing parameter on linking was not considered in this study. It would be useful to compare the separate and concurrent calibration methods using a bifactor model with a pseudo-guessing parameter, as this may be an alternative choice for calibrating dichotomously-scored multiple-choice items.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm, *Psychometrika, 46*, 443-459.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612.

Cai, L. (2017). *flexMIRT® version 3.51: Flexible multidimensional item analysis and test scoring* [Computer Software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods, 16*, 221-248.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *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., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Hanson, B. A., & Béguin, 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.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375-389.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics, 13*, 243-271.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics,* 38, 32-60.

Kang, T., & Petersen, N. S. (2011). Linking item parameters to a base scale. *Asia Pacific Education Review, 13,* 311-321.

Kim, S. (2006). A Comparative Study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131-143.

Kim, S., & Kolen, M. J. (2010). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19,* 357-381.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer.

Lee, W., & Ban, J. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23,* 23-48.

Lee, G., & Lee, W. (2016). Bi-factor MIRT observed-score equating for mixed format tests. *Applied Measurement in Education, 29,* 224-241.

Lee, G., Lee, W., Kolen, M. J., Park, I., Kim, D., & Yang, J. S. (2015). Bi-factor MIRT true-score equating for testlet-based tests. *Journal of Educational Evaluation, 28,* 681-700.

Li, Y. H., & Lissitz. R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*, 115-138.

Li, Y. H., & Lissitz. R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36,* 3-20.

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

Min, K. (2007). Evaluation of linking methods for multidimensional IRT calibrations. *Asia Pacific Education Review, 8,* 41-55.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37,* 357-373.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

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

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Tech. Rep. No. RR-09-03). Princeton, NJ: Educational Testing Service.

Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters.* (Unpublished doctoral dissertation). The University of Minnesota, Minneapolis, MN.

Woodruff. D. J., & Hanson, B. A. (1997). *Estimation of item response models using the EM algorithm for finite mixtures* (ACT Research Report Series 96-6). Iowa City, IA: ACT.

Woods, C. M., & Thissen. D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281-301.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*, 48-66.

Table 1

Mean and Standard Error of the Scaling and Translation Parameter Estimates for Separate Calibration and Those of the First Four Moments for Concurrent Calibration

			θ_0	θ_1	θ_2	θ_3
SC	CI = 20%	N = 2,000				
		Scaling	1.01 (.05)	1.01 (.12)	1.00 (.10)	1.01 (.13)
		Translation	-0.96 (.19)	0.03 (.28)	-0.46 (.30)	-0.97 (.27)
		N = 5,000				
		Scaling	1.01 (.03)	1.02 (.07)	0.99 (.05)	1.01 (.07)
		Translation	-0.94 (.11)	0.01 (.15)	-0.48 (.17)	-1.01 (.16)
	CI = 40%	N = 2,000				
		Scaling	1.01 (.04)	1.02 (.10)	1.00 (.08)	1.00 (.08)
		Translation	-0.93 (.12)	-0.01 (.22)	-0.50 (.21)	-1.01 (.21)
		N = 5,000				
		Scaling	1.01 (.02)	1.01 (.06)	0.99 (.05)	1.00 (.05)
		Translation	-0.92 (.09)	-0.02 (.14)	-0.51 (.13)	-1.02 (.13)
CC	CI = 20%	N = 2,000				
		Mean	1.07 (.05)	-0.07 (.07)	0.37 (.06)	0.86 (.07)
		SD	1.01 (.04)	1.06 (.06)	1.01 (.06)	1.05 (.06)
		Skewness	0.03 (.11)	0.15 (.22)	0.10 (.21)	0.27 (.13)
		Kurtosis	3.17 (.26)	3.60 (.55)	3.56 (.71)	3.50 (.29)
		N = 5,000				
		Mean	1.07 (.03)	-0.07 (.04)	0.38 (.04)	0.86 (.04)
		SD	1.01 (.02)	1.06 (.04)	1.01 (.03)	1.04 (.03)
		Skewness	0.04 (.08)	0.13 (.11)	0.12 (.13)	0.27 (.08)
		Kurtosis	3.10 (.18)	3.47 (.29)	3.44 (.37)	3.44 (.17)
	CI = 40%	N = 2,000				
		Mean	1.08 (.04)	-0.10 (.06)	0.36 (.06)	0.87 (.06)
		SD	1.00 (.04)	1.04 (.05)	1.00 (.06)	1.05 (.05)
		Skewness	0.03 (.10)	0.16 (.18)	0.13 (.17)	0.29 (.11)
		Kurtosis	3.16 (.25)	3.54 (.45)	3.45 (.48)	3.47 (.13)
		N = 5,000				
		Mean	1.08 (.03)	-0.10 (.03)	0.37 (.03)	0.88 (.04)
		SD	1.01 (.02)	1.04 (.04)	1.01 (.03)	1.05 (.03)
		Skewness	0.05 (.08)	0.14 (.10)	0.14 (.11)	0.29 (.07)
		Kurtosis	3.10 (.16)	3.46 (.24)	3.43 (.30)	3.43 (.15)

Notes. SC = separate calibration; CC = concurrent calibration; the values in the parentheses are the Monto Carlo standard errors.

Table 2

ABIAS, ASE, and ARMSE for Item Parameter Estimates

			a_0	a_1	a_2	a_3	d
ABIAS	CI = 20%	N = 2,000					
		Separate	0.02	0.01	0.01	0.01	0.01
		Concurrent	0.01	0.02	0.01	0.01	0.02
		N = 5,000					
		Separate	0.02	0.01	0.00	0.01	0.01
		Concurrent	0.01	0.01	0.01	0.01	0.02
	CI = 40%	N = 2,000					
		Separate	0.03	0.01	0.01	0.01	0.01
		Concurrent	0.01	0.01	0.01	0.01	0.01
		N = 5,000					
		Separate	0.02	0.01	0.00	0.01	0.01
		Concurrent	0.01	0.00	0.01	0.01	0.01
ASE	CI = 20%	N = 2,000					
		Separate	0.08	0.10	0.09	0.12	0.10
		Concurrent	0.08	0.08	0.08	0.10	0.09
		N = 5,000					
		Separate	0.05	0.06	0.06	0.07	0.06
		Concurrent	0.05	0.05	0.05	0.06	0.05
	CI = 40%	N = 2,000					
		Separate	0.08	0.09	0.09	0.11	0.09
		Concurrent	0.07	0.07	0.08	0.09	0.07
		N = 5,000					
		Separate	0.05	0.06	0.06	0.06	0.06
		Concurrent	0.04	0.04	0.05	0.05	0.05
ARMSE	CI = 20%	N = 2,000					
		Separate	0.09	0.10	0.09	0.13	0.10
		Concurrent	0.08	0.08	0.08	0.10	0.09
		N = 5,000					
		Separate	0.06	0.06	0.06	0.07	0.06
		Concurrent	0.05	0.05	0.05	0.06	0.06
	CI = 40%	N = 2,000					
		Separate	0.08	0.10	0.09	0.11	0.09
		Concurrent	0.07	0.07	0.08	0.09	0.08
		N = 5,000					
		Separate	0.05	0.06	0.06	0.06	0.06
		Concurrent	0.04	0.04	0.05	0.05	0.05

Table 3

ARMSE for Item Parameter Estimates across Common and Unique Items

			a_0	a_1	a_2	a_3	d
Common Items	CI = 20%	N = 2,000					
		Separate	0.08	0.09	0.07	0.10	0.07
		Concurrent	0.05	0.06	0.05	0.06	0.05
		N = 5,000					
		Separate	0.05	0.06	0.04	0.05	0.04
		Concurrent	0.03	0.04	0.03	0.04	0.03
	CI = 40%	N = 2,000					
		Separate	0.08	0.10	0.08	0.09	0.08
		Concurrent	0.05	0.06	0.06	0.05	0.05
		N = 5,000					
		Separate	0.05	0.06	0.05	0.05	0.05
		Concurrent	0.03	0.04	0.04	0.03	0.03
Unique Items	CI = 20%	N = 2,000					
		Separate	0.09	0.10	0.10	0.13	0.11
		Concurrent	0.08	0.08	0.09	0.11	0.10
		N = 5,000					
		Separate	0.06	0.06	0.06	0.07	0.07
		Concurrent	0.05	0.05	0.06	0.07	0.06
	CI = 40%	N = 2,000					
		Separate	0.09	0.09	0.09	0.12	0.10
		Concurrent	0.08	0.08	0.09	0.12	0.09
		N = 5,000					
		Separate	0.06	0.06	0.06	0.07	0.06
		Concurrent	0.05	0.05	0.06	0.07	0.06

Table 4

Bias, SE, and RMSE for Test Characteristic Surface (TCS) and Expected Observed-score Distribution (EOD) Criteria

			Uniform Weight			Normal Weight		
			Bias	SE	RMSE	Bias	SE	RMSE
TCS	CI = 20%	N = 2,000						
		Separate	0.11	0.64	0.65	0.11	0.53	0.54
		Concurrent	0.10	0.51	0.52	0.09	0.45	0.46
		N = 5,000						
		Separate	0.16	0.38	0.42	0.14	0.32	0.36
		Concurrent	0.07	0.30	0.31	0.08	0.28	0.30
	CI = 40%	N = 2,000						
		Separate	0.14	0.49	0.51	0.13	0.40	0.42
		Concurrent	0.07	0.42	0.43	0.06	0.37	0.38
		N = 5,000						
		Separate	0.17	0.32	0.37	0.14	0.26	0.30
		Concurrent	0.05	0.27	0.28	0.05	0.24	0.25
EOD	CI = 20%	N = 2,000						
		Separate	0.17	0.68	0.68	0.09	0.41	0.41
		Concurrent	0.13	0.54	0.54	0.07	0.35	0.35
		N = 5,000						
		Separate	0.16	0.42	0.42	0.11	0.27	0.27
		Concurrent	0.08	0.31	0.31	0.06	0.22	0.22
	CI = 40%	N = 2,000						
		Separate	0.15	0.53	0.53	0.10	0.32	0.32
		Concurrent	0.09	0.44	0.44	0.05	0.29	0.29
		N = 5,000						
		Separate	0.17	0.37	0.37	0.11	0.23	0.23
		Concurrent	0.06	0.28	0.28	0.04	0.18	0.18

Note. Bias, SE, and RMSE for the EOD criterion are the average over all possible score points.

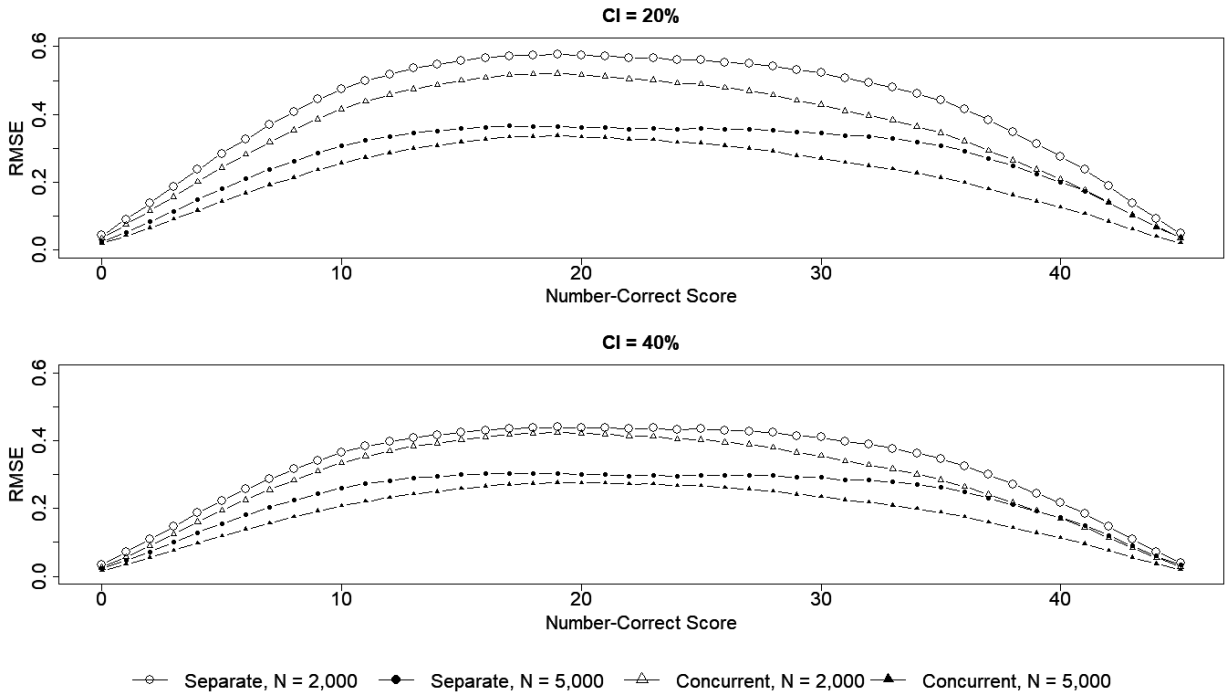


Figure 1. Values of conditional RMSE at each number-correct score for the EOD criterion.