Center for Advanced Studies in
Measurement and Assessment

CASMA Monograph

Number 2.4

Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 4)

Michael J. Kolen
Won-Chan Lee
(Editors)

December, 2016

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 4

This monograph, *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 4*), continues the work presented in Volumes 1-3 (Kolen & Lee, 2011, 2012, 2014). As stated in the Preface of the first volume,

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

Refer to the Preface for Volume 1 for more background on this research.

Volume 4 contains 9 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 9 address research on psychometric methods for mixed-format exams.

We thank, CASMA Psychometrician Jaime Malatesta, past graduate student Yujin Kang, and current graduate students Shichao Wang, Stella Kim, and Huan Liu 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
December, 2016
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.

Contents

Chapter 1: Introduction and Overview for Volume 4

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. It highlights some of the major findings, and 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 4

The research described in Volume 4 is closely related to research conducted in three previous volumes (Kolen & Lee, 2011, 2012, 2014). This chapter provides an overview of Volume 4 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 4 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 and 3 investigate various aspects of equating of mixed-format tests. Chapter 4 investigates the degree of interchangeability of FR items that are built to the same task model, with the idea that if the items can be used interchangeably then items built to the same task model might be used as common items in equating. Chapters 5 and 6 investigate estimation of classification consistency and accuracy for mixed-format tests. Chapter 7 investigates various multivariate generalizability theory designs for mixed-format tests. Chapter 8 investigates the development of score scales that have stable conditional standard errors of measurement along the score scale. Chapter 9 investigates the interval estimates of item response theory (IRT) proficiency estimates for mixed-format tests. As suggested by these studies taken as a whole, psychometric analyses for mixed-format tests often require complex extensions of methods used with single-format tests.

Chapter 2

In Chapter 2, the effect of various characteristics of common item sets on equating accuracy is investigated as an extension to studies by Lee, He, Hagge, Wang, and Kolen (2012), Wang and Kolen (2014), and Pak and Lee (2014) that were from chapters in previous monographs in this series. Item parameter estimates from various AP exams are used to form item pools. Forms are simulated from these item pools. This study concludes that equating is more accurate when (a) equal weights for MC and FR sections (i.e., equal contribution to composite scores) are used to equate with a mixed-format common item set, (b) mixed-format

common item sets, as opposed to MC-only common items, are used, and (c) the proportion of points associated with the common items is larger.

Chapter 3

In Chapter 3, the use of unidimensional IRT linking and equating with mixed-format tests is investigated. IRT separate, concurrent, and fixed calibration methods are considered along with true and observed score equating. A form of each of the three AP tests is split into two pseudo forms that contains common items. In addition, the examinee group taking each pseudo form is, based on demographic variables, divided into groups that differed in ability. The use of pseudo forms and pseudo groups allows for the study of the effect of variations in the common item sets and examinee group differences on equating accuracy. In addition, the pseudo forms design allows for the use of a single group criterion equating. A major finding is that concurrent calibration appears to be more accurate than separate or fixed calibration when only MC items are included in the common item set. When a mixed-format common item set is used, the calibration methods produced results that are similar in accuracy.

Chapter 4

Task models are used in the development of FR items for three AP exams. In Chapter 4 the extent to which items developed from the same task model can be used interchangeably is investigated. Methods used to detect unstable MC common items and differential item functioning are adapted and applied to the FR items built from the same task model. At least one task model per subject matter area was found to have items that behaved similarly across forms. The next step in this line of research is to investigate whether items built from the same task model can be used as common items to improve equating.

Chapter 5

In Chapter 5, classification consistency and accuracy for mixed-format AP tests are investigated. Results from various classical, unidimensional IRT, and multidimensional IRT procedures are compared. In general, the methods produce similar results. One interesting finding is that as data become more multidimensional, unidimensional and multidimensional IRT methods tend to produce different results.

Chapter 6

In Chapter 6, the effects of atypical score distributions on classification consistency and accuracy are compared. IRT item parameter estimates based on AP data are used as item

parameters for a simulation study. The atypical ability distributions investigated include distributions that are bimodal, bumpy, and that have structural zeros. The effects of atypical score distributions depended on the method used to estimate classification consistency and accuracy.

Chapter 7

In Chapter 7, multivariate generalizability theory is used to analyze data from AP mixed-format exams. Several multivariate designs are investigated, and the associated reliability estimates are compared. Based on the analyses, reliability estimates tend to be higher when the model considers errors associated with rater variability.

Chapter 8

In Chapter 8, three methods for constructing score scales that produce stable conditional standard errors of measurement for mixed-format tests are compared. The methods included are the arcsine transformation, the general variance stabilizing transformation, and a cubic transformation. The methods are applied to pseudo tests constructed from AP exams. It was found that all three methods stabilize the conditional standard errors of measurement.

Chapter 9

In Chapter 9, various methods for interval estimation of the IRT proficiency parameter for mixed-format tests are compared. The estimation methods investigated included maximum likelihood, method of moments, and Bayesian. The methods are examined through a simulation study based on mixed-format AP data. Results indicate that one of the maximum likelihood methods has more accurate coverage than other methods.

Discussion and Conclusions

This volume along with Volumes 1, 2, and 3 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.

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

Pak, S., & Lee, W. (2014). An investigation of performance of equating for mixed-format tests using only multiple-choice common items. 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.

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 Number 2.3).* Iowa City, IA: CASMA, The University of Iowa.

Chapter 2: Composition of Common Items for Equating with Mixed-Format Tests

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

Abstract

The growing popularity of mixed-format tests poses some issues in equating. Although a great deal of research has been done on equating, not much of the research has explored the effects of the characteristics of a common-item (CI) set on equating. The primary purposes of this study are to examine the effects on equating of various factors affecting the composition of a CI set, including (a) weights for multiple-choice (MC) and free-response (FR) section scores, (b) the proportion of CI scores relative to the total score, and (c) the composition of the CI set (i.e., a MC only CI set versus a CI set with both item formats). The general findings of this study include: (1) using equal weights for MC and FR sections produces more accurate equating results than using unequal weights for raw scores when a mixed-CI set is used; (2) the CI sets that include both item types tend to perform better than the MC-only CI sets; (3) better equating results tend to be associated with higher CI proportions; and (4) using 20% of the total test length as a conventional benchmark in a CI set seems reasonable.

Composition of Common Items for Equating with Mixed-Format Tests

Mixed-format tests have been gaining popularity in recent years and have been adopted by many testing programs. Containing both multiple-choice (MC) items and free-response (FR) items, mixed-format tests have an advantage in that they maximize benefits of each item type by using a combination of the two item formats. Specifically, MC items have strengths that include broader content coverage, efficiency of administration, and reliable and objective scoring. The principal advantage of FR items is in their ability to measure complex skills and in-depth knowledge.

In operational testing programs, equating is necessary because it allows scores from multiple forms of a test to be used interchangeably. The growing number of mixed-format tests has posed many issues in equating. Equating for mixed-format tests involves some practical and psychometric difficulties when a common-item (CI) set consists of MC items only, as is often the case in practice. A CI set is typically recommended to be proportionally representative of the total test in statistical and content specifications (Kolen & Brennan, 2014). The use of only one item type (e.g., MC items) in a CI set while the total test contains both MC and FR item types clearly violates the recommendation and, consequently, could lead to inaccurate equating results.

A few studies have been conducted to examine the effects of using an MC-only CI set and to explore conditions in which adequate levels of equating can be achieved. The factors that were found to be important with mixed-format equating include (a) correlation between MC and FR scores (Kirkpatrick, 2005; Lee, He, Hagge, Wang, & Kolen, 2012; Pak & Lee, 2014; Paek & Kim, 2007; Walker & Kim, 2009; Wang & Kolen, 2014), (b) similarity of the correlation between MC and FR scores for the old and new forms (Pak & Lee, 2014), (c) group ability differences (Cao, 2008; Kirkpatrick, 2005; Lee et al., 2012; Pak & Lee, 2014; Wang & Kolen, 2014), (d) similarity of group differences for MC and FR scores (Hagge & Kolen, 2011; Pak & Lee, 2014; Wang, 2013), (e) number of common items (Wang, 2013), (f) correlation between CI scores and total test scores (Wu, Huang, Huh, & Harris, 2009), (g) ratio of MC to FR score points (Paek & Kim, 2007; Tan, Kim, Paek, & Xiang, 2009), and (h) sample size (Pak & Lee, 2014). In sum, better equating results are expected as the correlation between MC and FR scores increases, sample size increases, the length of a CI set increases, the group difference becomes

smaller, the correlation level between MC and FR scores for old and new forms becomes more similar, and the level of group difference for MC and FR scores becomes more similar.

A limited number of studies have been conducted to evaluate the effects of CI set composition on equating with mixed-format tests (e.g., a CI set with MC items only versus a CI set with both MC and FR items). More specifically, Kirkpatrick (2005) found that the performance of a CI set with or without FR items largely depended on differential performance of examinees on sub-content areas. Mixed conclusions were reported by Hagge (2010), who explored the effects of a CI set composition in mixed-format test equating using a pseudo test form analysis. Recently, Wang and Kolen (2014) conducted a comprehensive study to examine the impact of different types of CI sets on mixed-format test equating under various conditions including item-type multidimensionality, group ability differences, and equating methods. Wang and Kolen (2014) concluded that when a group difference existed between old and new groups, a CI set with a mixture of MC and FR items produced a smaller equating error than a CI set with MC items only. Although FR items are seldom used as common items in practice, it is important to understand to what extent equating results may differ depending upon the use of FR items as CI under various conditions.

There are other factors that should be considered in equating practice that may impact equating results for mixed-format tests. First, composite scores on a mixed-format test often are constructed by weighted sums of MC and FR scores. Relative contributions of the MC and FR scores to the composite scores would affect the accuracy of equating results. The length of a CI set would also affect the magnitude of equating error. Regarding the number of CIs, a conventional rule of thumb is that a CI set should be at least 20% of the length of a total test (Kolen & Brennan, 2014). However, this suggestion is typically for tests consisting solely of MC items. Wang (2013) found that longer CI sets tend to lead to less equating error for mixed-format tests. Also, it was observed that the performance of the CI length condition tended to depend on the group difference and within-group difference for MC and FR sections. However, Wang (2013) only considered five levels for the combination of CI length and CI composition. Since the CI length condition was not fully crossed with other factors, the effects of CI length were not closely examined in a comprehensive way. Wang (2013) appears to be the only study that investigates how large the CI proportion needs to be to obtain adequate equating results for

mixed-format tests. Therefore, it is important to further explore the minimum level of CI proportion necessary for adequate equating under the mixed-format test settings.

The main purpose of this study is to investigate the effects on equating of (a) weights for MC and FR section scores, (b) the proportion of the points on a test that are included in a CI set, and (c) the composition of the CI set. In addition, this study examines how these factors interact with each other. A simulation study was conducted to address these research questions.

Method

Item Pool Generation

In order to construct item pools with realistic item parameters for the simulation study, two types of item pools – MC and FR – were created using the data from nine AP examinations administered in 2012, including Spanish Language and Culture, Spanish Literature, French Language and Culture, Italian Language and Culture, English Language and Culture, German Language and Culture, World History, Biology, and Environment Science. The selected examinations are all mixed-format tests that include both MC and FR items. The three parameter logistic model (Lord, 1980) and the graded response model (Samejima, 1997) were used to fit the MC and FR items, respectively, using flexMIRT (Cai, 2012). The estimated item parameters for the MC items were included in the MC item pool and the estimated FR item parameters were put in the FR item pool. There were 485 items in the MC item pool and 37 FR items in the FR item pool.

Test Form Construction

One new form and one old form were created from the item pool. These two forms were created in such a way that the distributions of the item parameters (in particular *a* and *c*) were as similar as possible. The total number of items on each form was fixed to 58. Among the 58 items, 50 items were MC items and 8 items were FR items. The FR section included two items each scored 0-10 and 6 items each scored 0-5. As a result, the MC and FR sections had the same number of score points (50 for each section) before any section weights were applied, such that unweighted composite scores ranged from 0 to 100. These two forms shared 20 MC items and four FR items scored 0-5 in common. Note that once the old and new forms were constructed, the same sets of item parameters were used to generate item responses across all study conditions (i.e., two forms are fixed).

Simulation Factors

MC and FR section weights. Three sets of integer section weights, (w_{MC}:w_{FR}), are considered: (1:3), (2:2), and (3:1). These sets of weights were selected to make the range of composite scores remain unchanged (0-200) under all weight conditions. Although operational AP examinations use non-integer weights to obtain composite scores, integer weights were selected in this study to simplify computations and to more efficiently deal with the magnitude and direction of various section weights.

Common item composition. Two types of CI set are explored: a CI set containing only MC items (abbreviated as MC-CI) and a CI set containing both MC and FR items (abbreviated as MX-CI). For a MX-CI set, the contribution of each item type to the total CI scores is equal (i.e., 50% for each item type).

Common item proportion. Various levels of the proportion of CI score points relative to composite score points are considered. Although the old and new forms were fixed across all study conditions, the CI proportion could be manipulated by removing a subset of common items from the CI set. That is, depending on the CI proportion condition, some common item were considered as non-common items.

The CI-proportion levels are confounded with the other two study factors of the CI composition and section weights. Four levels of unweighted CI proportion (.1, .2, .3, and .4) are first set on the metric of unweighted raw scores ranging from 0 to 100. After applying the weights to each of the section scores, the CI proportion relative to the composite score stays the same when the weights for MC and FR items are equal (see the middle rows with weights 2:2 in Table 1). However, when the weights for MC and FR are unequal, the CI proportion relative to the composite score changes. For example, when (3:1) weight is applied to the CI proportion of .1 with the MC-CI condition, the actual contribution of the CI scores to the composite scores becomes .15 because the unweighted raw score of 10 for the MC items is tripled with a weight of 3, which results in 30 CI score points out of 200 composite score points.

As a result of applying the section weights along with the two types of the CI composition, the range of CI proportion levels turns out to be .05-.60. Table 1 shows the CI-proportion conditions that were actually examined in this study, indicated by shaded cells. Note

that the levels of the CI-proportion factor are not fully crossed with the conditions of the other two study factors.

Correlation between MC and FR section scores. Regarding the correlation between MC and FR sections, Lee et al. (2012) used eleven levels of correlation from .50 to 1.0 with an increment of .05. Pak and Lee (2014) examined the effects of differential correlation on equating with the assumption that two forms could have a different correlation between MC and FR latent traits. In Pak and Lee's study (2014), nine levels of differential correlation were used with correlation values of .5, .8, and .95 for each form (i.e., 3 levels for each of two forms resulted in total 9 levels). Previous findings from Lee et al. (2012) and Pak and Lee (2014) indicate that as the correlation for both forms increases, the accuracy of equating results tends to improve. It is particularly noteworthy that a larger difference in the correlation between the MC and FR section scores for two forms (i.e., differential correlation) tends to result in a larger bias in equating.

Based on the previous findings, five levels of MC and FR correlation were considered in this study, including the differential correlation between old and new groups: (.7 & .7), (.7 & .95), (.8 & .8), (.95 & .7), and (.95 & .95). The first number indicates the correlation for the old group, and the second number is the correlation for the new group. As the correlation between two sections increases, the test can be viewed to become more unidimensional. When the disattenuated correlation is .8 or below, MC and FR items can be interpreted as measuring somewhat different constructs (Lee et al., 2012). In addition, previous research (Lee et al., 2012; Pak & Lee, 2014) found that, based on a DTM criterion, the minimum level of correlation level needed for adequate equating is .80, when the effect size for group differences is .2 or below.

Group ability difference (effect size). Under the common-item non-equivalent groups design, it is assumed that two groups differ in their ability levels. Group difference has been demonstrated to be an important factor affecting the accuracy of equating (Hagge & Kolen, 2011; Lee et al., 2012; Pak & Lee, 2014). Specifically, a larger group ability difference leads to a larger equating error. Lee et al. (2012) also observed that the frequency estimation method tended to perform worse than the chained equipercentile method when large group differences existed. According to the practical guidelines proposed by Lee et al. (2012), adequate equating results are expected to be achieved when the group ability difference is less than .2. Based on this suggestion, two levels of group ability difference were considered in this study: .1 as a small group difference and .3 as a large group difference, which is expressed as follows:

- Group difference = .1: $(\theta_{MC}, \theta_{FR})_{old} \sim BN(0,0,1,1,\rho)$ and $(\theta_{MC}, \theta_{FR})_{new} \sim BN(.1,.1,1,1,\rho)$
- Group difference = .3: $(\theta_{MC}, \theta_{FR})_{old} \sim BN(0,0,1,1,\rho)$ and $(\theta_{MC}, \theta_{FR})_{new} \sim BN(.3,.3,1,1,\rho)$

In the above equation, the pairs of MC and FR latent traits (θ_{MC}, θ_{FR}) are assumed to follow a bivariate normal distribution, $BN(\mu_{MC}, \mu_{FR}, \sigma^2_{MC}, \sigma^2_{FR}, \rho_{\theta_{MC}\theta_{FR}})$. Note that the new group was assumed to be more able than the old group and the differential level of group ability difference between MC and FR sections was not considered.

Equating and Simulation Procedures

Six equating methods were considered: unsmoothed frequency estimation (UnSm_FE), frequency estimation with log-linear presmoothing (PreSm_FE), frequency estimation with cubic-spline postsmoothing (PostSm_FE), unsmoothed chained equipercentile (UnSm_CE), chained equipercentile with log-linear presmoothing (PreSm_CE), and chained equipercentile with cubic-spline postsmoothing (PostSm_CE). The smoothing parameter for log-linear presmoothing was 6 for each marginal distribution and 1 for the cross-product, in keeping with common practice. For the cubic-spline postsmoothing, the smoothing parameter was .1, which has also been used by Lee et al. (2012). The bootstrap resampling procedure was conducted for 1,000 replications to obtain an estimate of the standard error of equating for cubic-spline postsmoothing. Synthetic population weights were 1 and 0 for the new and old groups, respectively. The reason for using this specific synthetic population weights is described in a later part of this section.

For all simulation conditions, the sample size was fixed at 3,000 for each of the old and new groups. The following is the simulation procedure used in this study:

1. Draw 3,000 random pairs of theta values from each of old and new group population distributions.
2. Generate item responses for each form using the item and person (theta) parameter values. Based on the simple-structure multidimensional IRT model, pairs of theta values were used to generate each of MC and FR item responses for each examinee.
3. Compute the composite and CI scores for each of the new and old forms.
4. Conduct equating based on the simulated data using the six equating methods listed previously.
5. Repeat the above steps 100 times.

Operationally, each examinee receives an integer grade level score ranging from 1 to 5. Since the AP grades are the primary score scale for AP, this study examines equating results for AP grades as well as raw scores. A conversion table that converts raw scores to AP grades was arbitrarily created for the old form. The cut scores corresponding to grade levels were determined based on the average observed distributions for four AP examinations so that the percent of examinees assigning to each grade was similar to that of the group who actually took the AP examinations. The resultant four cut scores were 91, 111, 130, and 148. The same cut scores were used for all study conditions. Equating results were evaluated in terms of unrounded raw and AP grade equated scores.

Criterion Equating Relationships

A single group design was used to establish the criterion equating relationships. The new-form population was assumed to take both forms, and traditional equipercentile equating was conducted based on the data obtained from the single population. A large sample (N=1,000,000) was drawn from the new-form population, $(\theta_{MC}, \theta_{FR})_{new} \sim BN(\mu_{MC}, \mu_{FR}, 1, 1, \rho_{\theta_{MC}\theta_{FR}})$ and item responses for the old and new forms were generated. Note that the new form population can change as study conditions change. Note also that the criterion equating relationships were determined based on the new form population, which was the basis for giving the full weight to the new group when synthetic populations are involved in equating. Thirty sets of criterion equating relationships were obtained; specifically, for two levels of group ability difference, five levels of correlations between MC and FR sections and three levels of MC and FR weights were obtained. The single group equipercentile equating is considered here to serve as an appropriate criterion because it is free from potential error due to the use of a CI set to represent the total test and two examinee groups that differ in ability. The single-group criterion used in this study is different from the criteria used in previous studies (Lee et al., 2012; Pak & Lee, 2014), which were based on the simple structure IRT model. A different equating criterion was employed in this study to examine whether or not simulation results depend on the choice of equating criterion.

Evaluation Criteria

Two evaluation criteria are employed in this study to evaluate equating results.

Raw score equating relationships. As summary statistics, weighted mean square error (MSE), weighted squared bias (SB), and weighted variance (VAR) were calculated to quantify the amount of error in equating. To prevent the overall results from being unduly affected by the equating results near the upper and lower ends of the score scale (where data are rarely observed), a valid observed score range was found for each replication. Then, only the score range where all study conditions have non-zero frequency was used for evaluation, which was the score range of 53-190. The overall statistics were obtained using a weighted sum of the conditional statistics over the score range of 53-190. Note that the weights used in this summation were defined as the relative frequency of each new form score point. The relative frequencies were adjusted based on the score range of 53- 190 such that the sum of the relative frequencies equaled one. Criterion indices were obtained using the following equations:

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

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

and

$$MSE(x) = SB(x) + VAR(x),$$

where e_x is the criterion equated score at raw score *x*; $\hat{e}_{xr}$ is an estimated equated score at raw score *x* on replication *r;* and w_x is the proportion of examinees with a new form raw score of *x*.

AP grade agreements. AP grade agreement statistics were computed by comparing the proportions for AP grades based on the criterion equating relationships with the proportions computed based on the estimated equating relationships. The differences between criterion classifications and the classifications based on the estimated equating results were indexed as MSE, SB, and VAR as expressed above, except that the parameter becomes the examinee proportion of each category. The evaluation criterion using AP grade agreements has not been examined in previous research.

Results

Results were evaluated with respect to raw score equating relationships and AP grade agreements. Although analyses were conducted for both raw scores and AP grade levels, the description of results focuses mainly on raw scores when a similar tendency was observed for AP grade levels, unless stated otherwise. In this section, the main effects for each study factor are described first, followed by the interaction effects between study factors.

Main Effects

Section weights. Tables 2 and 3 show the effects of section weights for raw scores. The results for AP grade agreements are presented in Tables 4 and 5. Note that due to the confounding conditions between the section-weight, CI-proportion, and CI-composition factors, two separate tables were created to discuss the results for the section weights—Table 2 (and Table 4) for the section-weight effects for the MC-CI condition and Table 3 (and Table 5) for the MX-CI condition. The results reported in these tables are the evaluation statistics aggregated over all study factors.

Table 2 shows the section-weight effects for the MC-CI condition in terms of a subset of the CI-proportion conditions: .1, .15, .2, and .3. There is a pair of section-weight conditions for each CI-proportion level. For example, the CI-proportion level of .1 (i.e., 20 score points) can be achieved by two different ways: 1) use 20 MC items in the CI set with a section weight of 1, or 2) use 10 MC items with a section weight of 2. In general, lower MSE and SB were observed for the conditions with a larger number of actual common-items than the conditions with a smaller number of common-items with a larger weight, except for the CI proportion of .30. It is conceivable that having more items in a CI set would increase the representativeness of the CI set to the total test, which cannot be attained by simply giving a larger weight to the common items.

When both MC and FR items are used as CIs, the CI proportion does not change regardless of the weight schemes because all items in a CI set are equally weighted. Thus, the effects of section weights could be observed more clearly with the MX-CI condition. Two evaluation criteria provide different results. In Table 3, the smallest MSE, SB, and VAR are always associated with the equal section-weight condition (i.e., 2:2). This suggests that when a CI set is composed of mixed item types, using equal numbers of score points for the MC and FR items in the CI set would be better than using unequal numbers of score points. The MSE and SB

values are largest for the 3:1 weight condition, indicating that assigning a larger weight to the MC common items produces less accurate equating results. However, in terms of the AP grade agreement statistics, different trends are observed. As shown in Table 5, equal section weight condition produces the largest errors among the three conditions. The other two unequal weight conditions provide similar results.

Common-item proportion. Figures 1 and 2 depict the changes in MSE, SB, and VAR for the MC-CI and MX-CI conditions, respectively, as a function of the CI proportions based on the raw-score equating results. The results reported in Figure 1 could be used to answer the question of what the minimum level of the CI proportion should be when a CI set consists of MC items only. Note that the AP operational equating involves MC-only CIs. Scree plots were created to visually inspect the minimum CI proportion to attain adequate equating results. As shown in the scree plots, as the CI proportion increases, both SB and VAR decrease, as expected. Note that the slight increase in error from the CI proportion of .4 to .45 in Figure 2 appears to be attributable to the confounded weight effects—that is, the CI proportion of .40 contains three MX-CI conditions, while the CI proportion of .45 involves one MC-CI condition.

Specifically, as seen in Figures 1 and 2, a practical benchmark of 20% CI proportion seems to be reasonable in that both plots exhibit an “elbow” occurring at .2. This pattern becomes more noticeable when the CI set contains both item formats. Even for the MC-CI condition, the amount of decrease in MSE and SB is less evident beyond the .2 range.

Common-item composition. Because the conditions of the CI composition were confounded with the conditions of the section weights and CI proportions, it was necessary to define comparable conditions to evaluate the CI composition effects. More specifically, the CI composition effects can only be investigated by comparing conditions that have the same section weights and CI proportion but different CI compositions.

As displayed in Tables 6 and 7, the MX-CI condition tends to perform better than the MC-CI condition in terms of all three evaluation statistics. The difference in MSE between the two CI-composition conditions is primarily due to the difference in SB. This finding is consistent with results reported by Wang and Kolen (2014).

Equating methods. The comparison between the FE and CE methods in Tables 8 and 9 suggests that the CE methods produce noticeably smaller SB values than the FE methods, which leads to smaller MSE values for the CE methods. The VAR values for the CE methods are

always larger than those for the FE methods. Similar findings have been observed in previous studies (Lee et al, 2012; Pak & Lee, 2014; Wang & Kolen, 2014). Note that the simulated data in this study were generated to have fairly large group differences in ability levels (i.e., .1 and .3), and thus the performance of the FE method in terms of bias was expected to be worse than the CE method.

Smoothing methods. As shown in Tables 8 and 9, two evaluation criteria lead to different conclusions. According to the raw-score equating results, the MSE values for the unsmoothed methods are larger than those for the two smoothing methods. The smallest VAR is generally found for the presmoothed methods and the largest is observed for the unsmoothed methods.

However, in terms of the AP grade agreement criterion (Table 9), the unsmoothed methods seem to perform slightly better than the two smoothing methods in terms of the MSE, primarily due to the smaller SB. This indicates that the amount of reduction in MSE by using smoothing methods depends on the criteria of interest.

Correlation between MC and FR traits. The effects of correlation between MC and FR latent trait variables (θs) are presented in Tables 10 and 11. Note that in the subsequent tables and discussion, "correlation" means the correlation between the MC and FR latent traits. In general, as correlation decreased, MSE, SB, and VAR tend to increase. In terms of differential correlation, the two evaluation criteria produce inconsistent results. The raw-score equating results are similar to previous study results (Pak & Lee, 2014), whereas the AP grade agreement results are somewhat different. In the previous study (Pak & Lee, 2014), it was found that larger differences in the correlation for two forms were associated with larger bias. This was true for the raw-score equating results, as shown in Table 10. The differential correlation between the MC and FR traits for the two forms (i.e., (.95&.7) and (.7&.95)) seems to cause a larger error, even compared to the smallest but equal correlation condition (i.e., (.7&.7)).

However, based on the AP grade agreement results displayed in Table 11, larger SB and MSE seem to be associated with the lower correlation for the new form. As a result, the smallest correlation for the new form with the largest correlation for the old form (.95&.7) produces an SB that is almost twice as large as the reversed correlation condition (.7&.95). Further research is needed to clarify why the two criteria produce different results.

Group difference (effect sizes). The effects of group difference are provided in Tables 12 and Table 13. The implication is straightforward. As the effect size gets larger, the magnitude of SB tends to become larger while the VAR remains almost constant over the two effect size conditions. Similar results were found in previous research (Lee et al., 2012; Pak & Lee, 2014; Wang & Kolen, 2014).

Interaction Effects

Common-item composition and correlation. The results for this section are summarized in Table 14. When there is a differential correlation between old-form and new-form groups (i.e., .7 & .95, .95 & .7), the MC-CI condition yields remarkably larger SB values than the MX-CI condition. This result clearly suggests that when the correlation between two sections is quite different for the two forms/groups, using both types of items in a CI set can reduce equating error substantially. For the MX-CI condition, the rank order of the MSE from smallest to largest is (.95 & .95), (.8 & .8), (.7 & .7), (.7 & .95), and (95 & .7), which suggests that equal correlation is preferred over differential correlation, even for the MX-CI condition.

For both CI-composition types, higher correlation always leads to lower error when correlation is equal for the two forms. Moreover, when both forms are almost unidimensional (i.e., .95 & .95), the MC-CI condition results in slightly smaller error than the MX-CI condition. This indicates that even when a full-length test contains both types of items, if all items measure a single (or similar) construct, using only MC items in a CI set produces results that are as accurate as or even better than using mixed item types in a CI set.

As shown in Table 15, the AP grade agreement results present somewhat different results. As mentioned earlier, smaller errors are generally found with the lower correlation conditions for the old form (i.e., (.7 & .95)) no matter which type of CI composition is employed. Consequently, a large amount of reduction in errors using the MX-CI can be achieved only for the differential correlation condition of (.95 & .7), which produces the largest errors across the five correlation conditions.

Common-item composition and group difference. Tables 16 and 17 present results for the interaction effects of group difference (effect size) and CI composition. As seen in the tables, the magnitude of the SB tends to increase as the group difference increases, regardless of the type of CI composition. Similar results have been observed by Wang (2013).

Common-item compositions and equating methods. As shown in Tables 18 and 19, the patterns of magnitude of error for various equating methods seem to be consistent no matter which type of CI composition is used. In general, the CE methods produce smaller MSE and SB than the FE methods for both CI-composition conditions. Also, the FE methods provide slightly smaller VAR than the CE methods.

Common-item proportion and correlation. The interaction effects between the CI proportion and correlation between MC and FR traits can be seen in Tables 20 and 21. As the CI proportion increases, the reduction in magnitudes of the SB and MSE becomes substantially larger for the equal correlation conditions (i.e., .7 & .7, .8 & .8, and .95 & .95) than for the differential correlation conditions (i.e., .7 & .95 and .95 & .7). This implies that when differential correlation exists, increasing the CI proportion may be a lot less effective compared to the cases with equal or similar correlation.

Regarding the AP grade agreement criterion (Table 21), other noteworthy observations can be made. When the CI proportion was very small (i.e., .05 or .10), the smallest MSE and SB values are associated with the correlation level of (.7 & .95). In addition, regardless of the CI proportion level, the correlation level of (.95 & .7) yields the largest MSE and SB among all five correlation levels. The reason for this finding is not clear.

Common-item proportion and group difference. Tables 22 and 23 show that increasing the CI proportion results in smaller MSE, SB, and VAR for both effect size levels (.1 and .3), which is not surprising given that better equating is usually expected when a longer CI set is used. More specifically, the amount of reduction in the MSE becomes remarkably larger for $ES = .3$ than for $ES = .1$ as the CI proportion increases, primarily due to the larger reduction in the SB. Therefore, when groups differ substantially, it is much more beneficial to make the CI set longer.

Common-item proportion and equating methods. As reported in Tables 24 and 25, the differences among equating methods become smaller as the CI proportion increases. For the FE methods, the largest MSE is associated with presmoothing when the CI proportion is small (i.e., .05 or .10); however, as the CI proportion increases, the presmoothing method outperforms the postsmoothing method in terms of the MSE values, especially for the raw-score equating results. The presmoothed CE method has the smallest MSE across all smoothing, equating methods, and CI proportion conditions, with a few exceptions in the AP grade results.

The two evaluation criteria provide somewhat inconsistent results with respect to the performance of the unsmoothed FE method. According to the raw-score equating results, the unsmoothed FE method leads to the largest MSE values over the CI-proportion range of .15 - .60. However, based on the AP grade agreement criterion, it tends to perform better than the other methods within this CI proportion range, although the reason for this is not clear at this point.

Common-item composition and common-item proportion. The interaction effects between the CI proportion and CI composition can be found in Tables 26 and 27. It is worth noting that the MX-CI condition tends to reduce more bias than the MC-CI condition as the CI proportion increases. This trend is more evident for the raw-score equating results (i.e., Table 26) than for the AP grade agreement results (i.e., Table 27). However, the amount of reduction in random error (VAR) is fairly consistent regardless of the CI composition conditions.

Summary

The current study is an extension of previous studies conducted by Lee et al. (2012), Wang and Kolen (2014), and Pak and Lee (2014). The main purpose of the study is to explore the effects of several CI-composition-related factors on equating for mixed-format tests. Conducting equating typically involves various practical considerations that might affect equating results. When a test is made up of a mixture of item formats, equating becomes much more complicated. Most importantly, in reality, practical constraints preclude the possibility of using all types of item formats in a CI set. The degree to which using a single item format in a CI set impacts equating accuracy needs to be evaluated. Also, a rule of thumb for determining the length of a CI set is to use 20% of the total test length. However, this conventional rule has not been studied much in order to examine whether it produces an acceptable level of accuracy in equating across various realistic testing conditions. Another factor that is often neglected in the literature is the use of weights for MC and FR sections. Since the aforementioned factors could potentially affect equating with mixed-format tests, it is important to study how each factor interacts with other factors and how much impact each factor has on equating results.

Section Weights

The results of this study suggest that using equal weights for MC and FR sections leads to better equating results than using unequal weights when a mixed-CI set is used for raw scores. When AP grade level is the score scale of interest, however, unequal conditions provide better

results. When a CI set is composed of MC items only, using more items with a smaller weight is a better strategy than using fewer items with a larger weight, assuming both cases produce the same score points.

Obviously, having more items in a CI set that differ in content and statistical characteristics would make the CI set more representative of the total test. Simply assigning a large weight to a CI set would not make it more representative.

Common-item Composition

Regarding the CI composition types, the mixed CI sets tend to perform better than the MC-only CI sets. This pattern becomes more obvious under unequal correlation conditions, which indicates that if two forms differ in correlation between two item formats, using both item types in a CI set helps reduce equating error substantially. In relation to group differences, more accurate equating results are associated with a smaller group differences regardless of CI composition conditions. The performance of equating and smoothing methods become more similar as the CI proportion increases, no matter which type of the CI composition is used. In summary, the CI composition does not seem to interact with the group effect size and equating methods. It is clear that as the CI proportion increases, a CI set with both types of item format results in more reduction in error compared to a CI with the MC items only. This implies that when a CI set is solely composed of MC items, it is necessary to use more CIs to achieve equating accuracy comparable with that for a mixed-format CI set.

Common-item Proportion

As depicted in Figures 1 and 2, better equating results tend to be associated with higher CI proportions. A conventional benchmark of using 20% of the total test length in a CI set seems reasonable in that the amount of decrease in error becomes less distinguishable beyond 20%. When a CI set contains both item formats, this tendency becomes more evident. Also, some interaction effects are observed for the CI proportion with the correlation between MC and FR traits, effect size, and equating methods. The decreases in amount of equating error are more evident for equal correlations than differential correlations as the CI proportion increases. This finding suggests that if the correlation between MC and FR traits is fairly different for the two forms, accurate equating cannot be anticipated even with a large CI proportion. The effect size also interacts with the CI proportion. When the group difference is large, increasing the CI proportion leads to more substantial reduction in equating error. That is, when the two groups are

quite different in their abilities, increasing the CI proportion can be very effective in reducing equating error. With regard to equating methods, a larger CI proportion tends to make the relative performance of various equating and smoothing methods less distinguishable. This suggests that when a CI proportion is not large enough, the choice of smoothing and equating methods could be more critical because different methods will likely result in quite different equating results.

Equating and Smoothing Methods

Comparisons between equating methods reveal that the CE methods produce less bias, and consequently less MSE, than the FE methods, possibly due to the large effect-size conditions employed in this study. Regarding the performance of the smoothing methods, the two evaluation criteria provide somewhat different results, and it seems very difficult to conclude which smoothing method performs better. In terms of raw-score equating, smoothing methods produce better results than unsmoothed methods primarily due to the smaller random error. The performances of presmoothing and postsmoothing methods are similar. Similar conclusions have been made in other studies (e.g., Lee et al., 2012).

Discussion

Some limitations in this study should be noted. First, the simulation was conducted under the simple structure IRT framework. The assumptions of the IRT framework might not be consistent with the assumptions made by the traditional equipercentile equating methods used in this study (i.e., FE and CE methods). If the assumptions are seriously violated, the results of this study might not properly reflect the performance of the equating methods. Also, only one sample size condition was examined. Although one might argue that the sample size of 3,000 is sufficiently large to generalize the study results to many practical equating settings, it would be interesting to explore how the results would change under other sample size conditions. Another limitation is that the ability difference between old and new groups is assumed to be equal for the MC and FR sections. This assumption might not be perfectly met in real situations. Readers can refer to Lee et al. (2012) and Pak and Lee (2014) for results obtained under different sample-size and effect-size conditions.

Structural Zeros

In the present study, the effects of section weights are investigated. Although it is common to use non-integer weights in practice, integer weights were adopted in this study for

the sake of simplicity. However, using integer weights caused “structural zeros” in score distributions. Structural zeros refer to the occurrence of zero frequency at some score points that are impossible to attain. For example, with the (2:2) weights, examinees can never earn an odd number composite score point because every score point is multiplied by two. In addition, when a CI set is composed of MC items only, only one out of every three score points can be observed for the CI scores if the (3:1) weights are applied because the CI set solely contains MC items and all the items are triple-weighted. These structural zeros would probably impact the equating results, particularly when a presmoothing method is used because presmoothing produces some positive frequency values to score points with structural zeros.

Several approaches to handling structural zeros can be investigated in a future study. One possible way would be to conduct smoothing first and then apply weights. In this case, the structural zeros would remain in the score distribution even after the section weights are applied. Another possible method would be to assign zero frequencies to the structural zero points intentionally after smoothing; however, this method might, to some extent, lose the moment preservation property of smoothing methods.

References

Cai, L. (2012). *flexMIRT* (Version 1.88) [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

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

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.

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1). (CASMA Monograph Number 2.1)* (pp. 95-135). Iowa City, IA: CASMA, The University of Iowa.

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

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

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

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

Paek, I., & Kim, S. (2007, April). *Empirical investigation of alternatives for assessing scoring consistency on constructed response items in mixed format tests.* Paper presented at the 2007 annual meeting of the American Educational Research Association, Chicago, IL.

Pak, S., & Lee, W. (2014). An investigation of performance of equating for mixed-format tests using only multiple-choice common items. 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.

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

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

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

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

Wang, W., & Kelen, 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 Number 2.3).* Iowa City, IA: CASMA, The University of Iowa.

Wu, N., Huang, C., Huh, N., & Harris, D. (2009). *Robustness in using multiple-choice items as an external anchor for constructed-response test equating.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Diego, CA.

Table 1

Study Conditions of CI proportion, CI composition, and Section Weights

MC/FR Weights	CI Composition	CI Proportion							
		.05	.10	.15	.20	.30	.40	.45	.60
1:3	MC-CI	■	■	■	■				
	MX-CI		■		■	■	■		
2:2	MC-CI		■		■	■	■		
	MX-CI		■		■	■	■		
3:1	MC-CI			■		■		■	■
	MX-CI		■		■	■	■		

Note. Shaded cells indicate the conditions considered in this study.

Table 2

Aggregated Summary Statistics for Section Weights with CI Composition (MC-CI condition only)

CI Proportion (Score Points)		MC / FR Weights		
		1:3	2:2	3:1
Score Points = 20		20 MC items × Weight of 1	10 MC items × Weight of 2	
.10 (20)	MSE	5.47340	9.34795	
	SB	4.33455	8.22080	
	VAR	1.13898	1.12720	
Score Points = 30		30 MC items × Weight of 1		10 MC items × Weight of 3
.15 (30)	MSE	4.20593		7.23177
	SB	3.16230		6.27740
	VAR	1.04365		.95432
Score Points = 40		40 MC items × Weight of 1	20 MC items × Weight of 2	
.20 (40)	MSE	3.76198	5.24608	
	SB	2.73967	4.34698	
	VAR	1.02228	.89907	
Score Points = 60			30 MC items × Weight of 2	20 MC items × Weight of 3
.30 (60)	MSE		4.16233	3.21677
	SB		3.37803	2.58197
	VAR		.78435	.63475

Note. Shaded cells indicate the conditions not considered in this study.

Table 3

Aggregated Summary Statistics for Section Weights with CI Composition (MX-CI condition only)

	MC / FR Weights		
	1:3	2:2	3:1
MSE	2.97699	2.84893	3.42900
SB	2.18065	2.11223	2.65904
VAR	0.79635	0.73671	0.76999

Table 4

Aggregated AP Grade Agreements for Section Weights with CI Composition (MC-CI condition only)

CI Proportion (Score Points)		MC / FR Weights		
		1:3	2:2	3:1
Score Points = 20		20 MC items × Weight of 1	10 MC items × Weight of 2	
.10 (20)	MSE	.00032	.00073	
	SB	.00023	.00061	
	VAR	.00009	.00012	
Score Points = 30		30 MC items × Weight of 1		10 MC items × Weight of 3
.15 (30)	MSE	.00027		.00036
	SB	.00018		.00028
	VAR	.00008		.00008
Score Points = 40		40 MC items × Weight of 1	20 MC items × Weight of 2	
.20 (40)	MSE	.00027	.00045	
	SB	.00018	.00033	
	VAR	.00009	.00012	
Score Points = 60			30 MC items × Weight of 2	20 MC items × Weight of 3
.30 (60)	MSE		.00041	.00018
	SB		.00029	.00010
	VAR		.00012	.00008

Note. Shaded cells indicate the conditions not considered in this study.

Table 5

Aggregated AP Grade Agreements for Section Weights with CI Composition (MX-CI condition only)

	MC / FR Weights		
	1:3	2:2	3:1
MSE	.00020	.00032	.00020
SB	.00011	.00021	.00011
VAR	.00008	.00012	.00008

Table 6

Aggregated Summary Statistics for CI Composition

	CI Composition	
	MC-CI	MX-CI
MSE	4.99199	3.15269
SB	4.09116	2.36688
VAR	.90085	.78584

Note. Only comparable sets were selected.

Table 7

Aggregated AP Grade Agreements for CI Composition

	CI Composition	
	MC-CI	MX-CI
MSE	.00039	.00027
SB	.00029	.00017
VAR	.00009	.00008

Note. Only comparable sets were selected.

Table 8

Aggregated Summary Statistics for Equating Methods

	Equating Methods					
	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
MSE	5.45648	2.89510	5.36471	2.40723	5.35380	2.69446
SB	4.58760	1.75015	4.82556	1.72773	4.62750	1.74417
VAR	0.86889	1.14501	0.53917	0.67951	0.72628	0.95035

Table 9

Aggregated AP Grade Agreements for Equating Methods

	Equating Methods					
	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
MSE	.00031	.00022	.00036	.00023	.00033	.00024
SB	.00023	.00010	.00028	.00013	.00026	.00013
VAR	.00008	.00012	.00007	.00009	.00006	.00009

Table 10

Aggregated Summary Statistics for Correlation

	Correlation				
	(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
MSE	3.86800	4.77440	3.32706	5.50030	2.67340
SB	3.02239	3.92512	2.50746	4.69485	1.90243
VAR	0.84558	0.84935	0.81962	0.80548	0.77097

Table 11

Aggregated AP Grade Agreements for Correlation

	Correlation				
	(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
MSE	.00027	.00023	.00026	.00047	.00022
SB	.00018	.00013	.00017	.00038	.00013
VAR	.00009	.00007	.00007	.00010	.00009

Table 12

Aggregated Summary Statistics for Effect Size

	Effect Size	
	ES = .1	ES = .3
MSE	2.10539	5.95187
SB	1.28410	5.13681
VAR	.82133	.81507

Table 13

Aggregated AP Grade Agreements for Effect Size

	Effect Size	
	ES = .1	ES = .3
MSE	.00018	.00038
SB	.00009	.00029
VAR	.00008	.00009

Table 14

Aggregated Summary Statistics for Interaction of CI Composition and Correlation

CI Composition		Correlation				
		(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
MC-CI	MSE	4.67934	6.31856	3.74167	7.53465	2.58723
	SB	3.74179	5.39209	2.86385	6.69581	1.82445
	VAR	.93749	.92660	.87785	.83885	.76279
MX-CI	MSE	3.05665	3.23023	2.91246	3.46595	2.75956
	SB	2.30299	2.45815	2.15108	2.69390	1.98040
	VAR	.75368	.77210	.76139	.77210	.77915

Note. Only comparable sets of conditions were selected.

Table 15

Aggregated AP Grade Agreements AP Grade Agreements for Interaction of CI Composition and Correlation

CI Composition		Correlation				
		(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
MC-CI	MSE	.00037	.00026	.00032	.00078	.00023
	SB	.00026	.00016	.00021	.00067	.00013
	VAR	.00009	.00007	.00007	.00010	.00009
MX-CI	MSE	.00028	.00022	.00027	.00035	.00025
	SB	.00018	.00012	.00017	.00025	.00015
	VAR	.00008	.00008	.00008	.00008	.00008

Note. Only comparable sets of conditions were selected.

Table 16

Aggregated Summary Statistics for Interaction of CI Composition and Effect Size

CI Composition		Effect Size	
		ES = .1	ES = .3
MC-CI	MSE	2.79480	7.14978
	SB	1.92388	6.28332
	VAR	.87094	.86649
MX-CI	MSE	1.41598	4.75396
	SB	.64431	3.99029
	VAR	.77172	.76365

Note. Only comparable sets of conditions were selected.

Table 17

Aggregated AP Grade Agreements for Interaction of CI Composition and Effect Size

CI Composition		Effect Size	
		ES = .1	ES = .3
MC-CI	MSE	.00026	.00052
	SB	.00015	.00042
	VAR	.00009	.00009
MX-CI	MSE	.00017	.00038
	SB	.00007	.00028
	VAR	.00008	.00009

Note. Only comparable sets of conditions were selected.

Table 18

Aggregated Summary Statistics for Interaction of CI Composition and Equating Methods

CI Composition		Equating Methods					
		UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
MC-CI	MSE	6.62694	3.65353	6.46853	3.15597	6.48525	3.44351
	SB	5.70948	2.45080	5.88761	2.40692	5.71800	2.44879
	VAR	.91752	1.20274	.58098	.74905	.76722	.99479
MX-CI	MSE	4.28601	2.13667	4.26088	1.65849	4.22235	1.94542
	SB	3.46572	1.04951	3.76352	1.04854	3.53699	1.03955
	VAR	.82026	1.08728	.49736	.60998	.68533	.90591

Note. Only comparable sets of conditions were selected.

Table 19

Aggregated AP Grade Agreements for Interaction of CI Composition and Equating Methods

CI Composition		Equating Methods					
		UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
MC-CI	MSE	.00041	.00030	.00051	.00032	.00046	.00034
	SB	.00033	.00017	.00042	.00022	.00038	.00021
	VAR	.00006	.00012	.00008	.00011	.00006	.00011
MX-CI	MSE	.00029	.00021	.00035	.00022	.00034	.00024
	SB	.00021	.00008	.00027	.00012	.00026	.00012
	VAR	.00008	.00012	.00007	.00009	.00006	.00009

Note. Only comparable sets of conditions were selected.

Table 20

Aggregated Summary Statistics for Interaction of CI Proportion and Correlation

CI Proportion		Correlation				
		(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
.05	MSE	10.43683	9.23325	8.73117	11.62625	6.43000
	SB	9.02200	7.77717	7.43392	10.40733	5.21975
	VAR	1.41492	1.45633	1.29742	1.21925	1.21008
.10	MSE	7.16375	7.94723	6.48325	9.05218	5.45850
	SB	6.06165	6.79692	5.40790	7.89822	4.41458
	VAR	1.10208	1.1504	1.07537	1.15400	1.04393
.15	MSE	5.81396	6.47125	4.77133	7.91404	3.62367
	SB	4.73954	5.40758	3.72863	6.94592	2.77758
	VAR	1.07433	1.06367	1.04267	0.96804	0.84621
.20	MSE	3.29515	4.1447	2.67903	4.67668	2.14045
	SB	2.37762	3.25433	1.81612	3.85045	1.32083
	VAR	0.91753	0.89043	0.86290	0.82623	0.81962
.30	MSE	2.10657	3.22130	1.85092	3.67765	1.48800
	SB	1.41495	2.52890	1.15428	3.01378	0.81260
	VAR	0.69163	0.69247	0.69667	0.66393	0.67535
.40	MSE	1.35996	2.30769	1.12815	2.50373	0.92325
	SB	0.77527	1.73494	0.55550	1.95181	0.38581
	VAR	0.58465	0.57279	0.57269	0.55194	0.53748
.45	MSE	1.49675	3.43942	1.20433	3.97258	0.84050
	SB	1.02958	2.95150	0.74350	3.54283	0.40017
	VAR	0.46708	0.48833	0.46092	0.42983	0.44008
.60	MSE	1.00325	3.17333	0.79275	3.53283	0.51592
	SB	0.63458	2.81858	0.43092	3.21500	0.19983
	VAR	0.36850	0.35475	0.36183	0.31775	0.31625

Table 21

Aggregated AP Grade Agreements for Interaction of CI Proportion and Correlation

CI Proportion		Correlation				
		(.7 & .7)	(.7 & .95)	(.8 & .8)	(.95 & .7)	(.95 & .95)
.05	MSE	.00041	.00032	.00047	.00085	.00037
	SB	.00032	.00023	.00038	.00076	.00029
	VAR	.00009	.00008	.00008	.00009	.00008
.10	MSE	.00045	.00034	.00043	.00075	.00036
	SB	.00035	.00024	.00033	.00065	.00027
	VAR	.00009	.00007	.00007	.00010	.00009
.15	MSE	.00027	.00027	.00027	.00054	.00022
	SB	.00019	.00019	.00018	.00046	.00013
	VAR	.00008	.00008	.00008	.00009	.00008
.20	MSE	.00028	.00021	.00025	.00045	.00020
	SB	.00018	.00011	.00015	.00035	.00010
	VAR	.00009	.00008	.00010	.00010	.00008
.30	MSE	.00020	.00017	.00020	.00034	.00016
	SB	.00011	.00008	.00010	.00025	.00007
	VAR	.00012	.00011	.00012	.00012	.00012
.40	MSE	.00019	.00015	.00017	.00033	.00016
	SB	.00010	.00006	.00008	.00023	.00007
	VAR	.00012	.00011	.00011	.00011	.00012
.45	MSE	.00012	.00018	.00010	.00023	.00009
	SB	.00004	.00010	.00003	.00016	.00002
	VAR	.00007	.00008	.00007	.00008	.00008
.60	MSE	.00010	.00017	.00009	.00020	.00009
	SB	.00003	.00010	.00002	.00012	.00001
	VAR	.00007	.00007	.00007	.00008	.00007

Table 22

Aggregated Summary Statistics for Interaction of CI Proportion and Effect Size

CI Proportion		Effect Size	
		ES = .1	ES = .3
.05	MSE	3.78507	14.79793
	SB	2.50027	13.44380
	VAR	1.28510	1.35410
.10	MSE	3.36511	11.37686
	SB	1.94994	10.28177
	VAR	1.11520	1.09511
.15	MSE	2.77858	8.65912
	SB	1.77855	7.66115
	VAR	1.00002	.99795
.20	MSE	1.94648	4.82793
	SB	1.07529	3.97245
	VAR	.87123	.85546
.30	MSE	1.61157	3.32620
	SB	.92293	3.64688
	VAR	.68869	.67933
.40	MSE	1.22561	2.06350
	SB	.66369	1.49764
	VAR	.56192	.56590
.45	MSE	1.66237	2.71907
	SB	1.20267	2.26437
	VAR	.45973	.45477
.60	MSE	1.50647	2.10077
	SB	1.16270	1.75687
	VAR	.34373	.34390

Table 23

Aggregated AP Grade Agreements for Interaction of CI Proportion and Effect Size

CI Proportion		Effect Size	
		ES = .1	ES = .3
.05	MSE	.00021	.00076
	SB	.00012	.00067
	VAR	.00009	.00009
.10	MSE	.00023	.00070
	SB	.00013	.00060
	VAR	.00009	.00009
.15	MSE	.00017	.00045
	SB	.00009	.00037
	VAR	.00008	.00009
.20	MSE	.00018	.00037
	SB	.00008	.00027
	VAR	.00009	.00009
.30	MSE	.00016	.00027
	SB	.00007	.00018
	VAR	.00012	.00013
.40	MSE	.00016	.00024
	SB	.00007	.00014
	VAR	.00012	.00011
.45	MSE	.00012	.00016
	SB	.00005	.00009
	VAR	.00007	.00008
.60	MSE	.00012	.00013
	SB	.00005	.00006
	VAR	.00007	.00007

Table 24

Aggregated Summary Statistics for Interaction of CI Proportion and Equating Methods

CI Proportion		Equating Methods					
		UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
.05	MSE	13.78890	5.09940	13.81600	4.72060	13.53260	4.79150
	SB	12.47560	3.37640	12.81180	3.36090	12.43580	3.37170
	VAR	1.31360	1.72300	1.00430	1.35990	1.09690	1.41990
.10	MSE	10.08264	4.59758	10.28492	4.09320	9.92802	4.33954
	SB	8.96962	3.11416	9.49210	3.03600	8.98914	3.09410
	VAR	1.11298	1.48354	.79284	1.05722	.93888	1.24548
.15	MSE	7.80515	3.98465	7.71615	3.48885	7.58940	3.72890
	SB	6.76955	2.64125	6.98815	2.55020	6.73435	2.63560
	VAR	1.03565	1.34335	.72810	.93855	.85510	1.09315
.20	MSE	4.59952	2.50414	4.43780	1.98520	4.49890	2.29766
	SB	3.68806	1.29168	3.86490	1.27374	3.73544	1.28940
	VAR	.91146	1.21258	.57284	.71146	.76342	1.00830
.30	MSE	3.12348	2.11926	2.93230	1.61802	3.07172	1.94854
	SB	2.36640	1.12832	2.51844	1.13080	2.43882	1.12664
	VAR	.75702	.99106	.41390	.48726	.63284	.82198
.40	MSE	2.00140	1.57458	1.78620	1.09640	1.97390	1.43485
	SB	1.35455	.72780	1.48558	.75512	1.43392	.72702
	VAR	.64688	.84678	.30068	.34132	.53985	.70795
.45	MSE	2.51240	2.14260	2.32040	1.73140	2.44000	1.99750
	SB	1.97210	1.44390	2.08310	1.45910	2.00150	1.44140
	VAR	.54070	.69860	.23730	.27220	.43860	.55610
.60	MSE	2.01000	1.86790	1.76440	1.47620	1.95100	1.75220
	SB	1.57690	1.31890	1.62280	1.32190	1.60120	1.31700
	VAR	.43290	.54890	.14160	.15410	.34990	.43550

Table 25

Aggregated AP Grade Agreements for Interaction of CI Proportion and Equating Methods

CI Proportion		Equating Methods					
		UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
.05	MSE	.00065	.00031	.00070	.00030	.00064	.00030
	SB	.00058	.00019	.00063	.00019	.00058	.00019
	VAR	.00006	.00013	.00006	.00011	.00007	.00011
.10	MSE	.00055	.00031	.00065	.00034	.00059	.00035
	SB	.00046	.00018	.00057	.00023	.00052	.00023
	VAR	.00006	.00012	.00008	.00011	.00006	.00011
.15	MSE	.00039	.00024	.00041	.00023	.00037	.00023
	SB	.00032	.00014	.00035	.00013	.00031	.00013
	VAR	.00007	.00011	.00006	.00010	.00006	.00010
.20	MSE	.00030	.00022	.00035	.00022	.00032	.00024
	SB	.00022	.00009	.00027	.00012	.00025	.00012
	VAR	.00006	.00013	.00006	.00010	.00007	.00011
.30	MSE	.00021	.00019	.00026	.00019	.00024	.00021
	SB	.00013	.00007	.00018	.00010	.00017	.00010
	VAR	.00010	.00015	.00010	.00012	.00010	.00014
.40	MSE	.00018	.00018	.00023	.00019	.00021	.00021
	SB	.00010	.00006	.00015	.00009	.00014	.00009
	VAR	.00010	.00014	.00010	.00013	.00008	.00014
.45	MSE	.00014	.00016	.00015	.00013	.00014	.00015
	SB	.00007	.00006	.00008	.00006	.00008	.00006
	VAR	.00006	.00010	.00006	.00007	.00006	.00009
.60	MSE	.00013	.00014	.00012	.00012	.00013	.00013
	SB	.00006	.00005	.00006	.00005	.00006	.00005
	VAR	.00006	.00007	.00006	.00007	.00006	.00007

Table 26

Aggregated Summary Statistics for Interaction of CI Proportion and CI Composition

CI Composition		CI Proportion			
		.10	.20	.30	.40
MC-CI	MSE	7.41068	4.50403	3.68955	3.73542
	SB	6.27768	3.54333	2.98000	3.03615
	VAR	1.13309	.96068	.70955	.69933
MX-CI	MSE	6.58565	2.44713	1.56213	.87902
	SB	5.51815	1.65415	.92209	.37935
	VAR	1.06749	.79303	.64008	.49967

Note. Only comparable sets of conditions were selected.

Table 27

Aggregated AP Grade Agreements for Interaction of CI Proportion and CI Composition

CI Composition		CI Proportion			
		.10	.20	.30	.40
MC-CI	MSE	.00052	.00036	.00029	.00039
	SB	.00042	.00026	.00019	.00027
	VAR	.00009	.00009	.00012	.00012
MX-CI	MSE	.00044	.00024	.00018	.00021
	SB	.00034	.00014	.00008	.00010
	VAR	.00008	.00009	.00011	.00012

Note. Only comparable sets of conditions were selected.

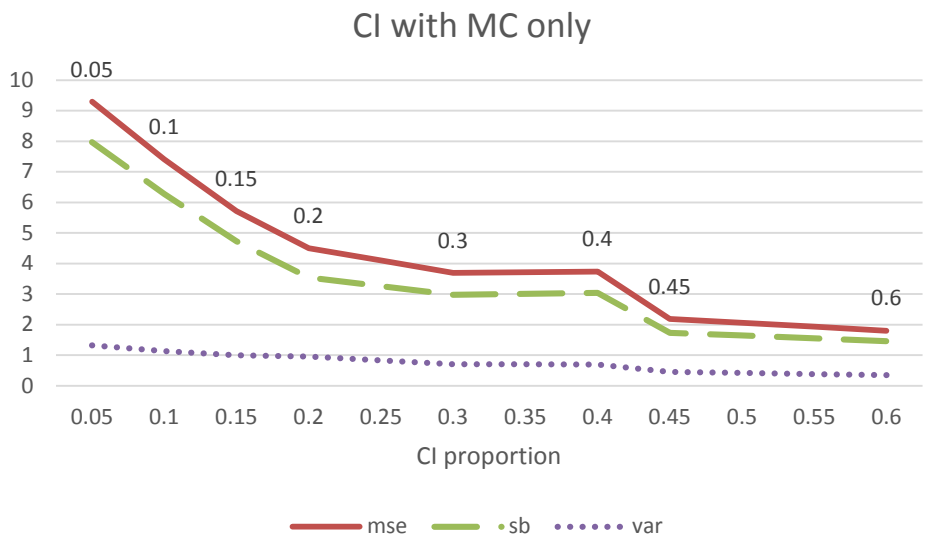

Figure 1. Equating error with respect to CI proportion (MC-CI condition only).

Note. Only the data points with labels above the solid line are actually observed.

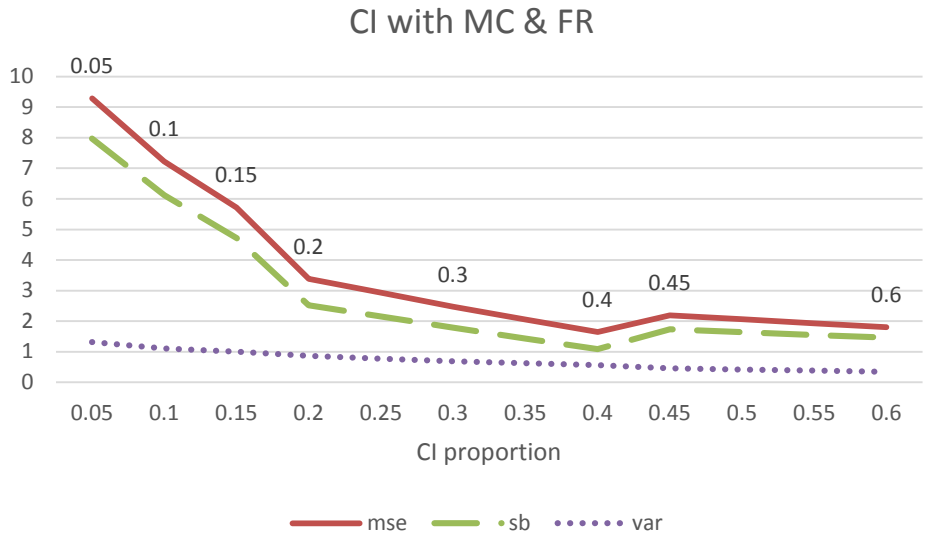

Figure 2. Equating error with respect to CI proportion (MX-CI condition only).

Note. Only the data points with labels above the solid line are actually observed.

Chapter 3: Comparison of IRT Linking and Equating Methods with Mixed-Format Tests

Yujin Kang and Won-Chan Lee
The University of Iowa, Iowa City, IA

Abstract

The purpose of this study is to investigate the robustness of unidimensional item response theory linking and equating methods to format effects with mixed-format tests. The linking methods considered are separate, concurrent, and fixed parameter calibration, and equating methods are observed score and true score equating. This research consists of two studies. The first study involves intact forms and groups, while the second study involves pseudo forms and groups with an equipercentile criterion. In the first study, three linking methods performed quite differently when format effects were large. It was found in the second study that concurrent calibration was more robust to format effects. The other two linking methods were similar to concurrent calibration in amount of equating error when both multiple-choice and free-response items were included in a common-item set. In both studies, the performance of the two equating methods was similar.

Comparison of IRT Linking and Equating Methods with Mixed-Format Tests

The purpose of this study is to examine the robustness of unidimensional IRT linking and equating methods to format effects. Format effect can exist in a mixed-format test consisting of a mixture of various item types—for example, multiple-choice (MC) and free-response (FR) items—when each item type represents a different construct (Kim & Kolen, 2006; Traub, 1993). In this case, it would be reasonable to conclude that multidimensionality exists for the test. Therefore, when IRT is applied to mixed-format tests with format effects, it has been argued that multidimensional IRT models should be used for the tests (Cao, 2008; Kolen & Lee, 2011). However, in practice, unidimensional IRT models might be preferred because it is easier to implement them in practice than multidimensional IRT models. This study is intended to investigate the performance of various unidimensional IRT linking and equating methods when applied to data with various degrees of format effects.

Background Information

IRT Linking and Equating

Equating is a statistical process to adjust scores on test forms so that the scores on different forms can be used interchangeably (Kolen & Brennan, 2014). Before collecting data for equating, an equating design should be chosen. This study is based on the common-item nonequivalent groups (CINEG) design. In the CINEG design, different groups of examinees take different test forms with a set of common items (CIs). Equating is conducted through the common item (CI) set. Among various equating methods for the CINEG design, IRT methods are considered in this study.

Equating using IRT methods typically consists of three steps (Cao, 2008; Kolen & Brennan, 2014). First, item and ability parameters are estimated using IRT models, which is often called “calibration.” Second, the parameter estimates of the two forms to be equated are transformed to be on the same scale, which is referred to as “scale transformation.” Third, equating typically is conducted on raw scores first, and then the resulting raw-score equivalents are converted to scale scores. The first and second steps can be combined for some procedures, and the term “linking” is used here to refer to the calibration and scale transformation combined together.

There are three types of IRT linking methods that are frequently discussed in the literature. First, the separate calibration with scale transformation (SEP) method involves

estimating parameters of each form separately, and then conducting scale transformation using a method such as the Stocking and Lord method (Stocking & Lord, 1983). Second, the multiple-group concurrent calibration (CON) method (Lord, 1980; Wingersky & Lord, 1984) estimates the parameters of two forms simultaneously based on the combined data. Those items that are not shared by two groups do not have item responses and they are treated as "not reached" items in the calibration process. Third, fixed-parameter calibration (FIX; Kim, 2006; Li, Tam, & Tompkins, 2004) estimates the parameters for Form Y (old or base form) first. Then, the estimated parameters for Form Y are fixed at their values when estimating the parameters for Form X (new form) so that all the parameter estimates will be on the Form Y scale.

With respect to IRT equating, two methods are often considered. True score equating (Lord & Wingersky, 1984) considers that true scores on two forms are equivalent if they correspond to the same value of the ability parameter. The final equating relationship is obtained by replacing true scores by observed scores. By contrast, observed score equating (Lord & Wingersky, 1984) uses estimated observed score distributions for the two forms based on an IRT model. Then, traditional equipercentile equating is conducted on the estimated distributions.

Previous Studies on the Robustness to Format Effects

There exist several studies on the robustness of unidimensional IRT linking methods to format effects. Kim and Kolen (2006) conducted a simulation study to compare the robustness of SEP and CON. They found that CON worked slightly better than SEP. Cao (2008) conducted a simulation study focusing on CON, and found that CON was not robust when a group difference was large and CIs were not format-representative. Some studies investigated the robustness of unidimensional IRT equating methods to format effects, in terms of equity properties (Andrews, 2011; He & Kolen, 2011; Wolf, 2013). These studies found, in general, that the performance of the true score and observed score equating methods was similar when used with SEP (Andrews, 2011; He & Kolen, 2011) and CON (Wolf, 2013).

There are several limitations of the previous studies. First, most studies focused on the robustness of either unidimensional IRT linking or equating methods only, and no studies have examined the linking and equating methods simultaneously and their interactions. Second, previous studies used one or two linking methods, and there is very little in the literature that compares all three linking methods. Third, all the studies discussed previously were based on simulations, and very little is known how these methods perform with real mixed-format data.

Research Objectives

The primary purpose of the present study is to examine the robustness of three unidimensional IRT linking methods, including SEP, CON, and FIX, in conjunction with two unidimensional IRT equating methods. Several real mixed-format datasets with various levels of format effects are used. This research consists of two studies. The first study employs actual intact forms and groups for analysis, while the second study involves pseudo forms and groups with an evaluation criterion.

Study I: Intact Forms and Groups

In the first study, real data with varying format effects are analyzed. This study uses actual data without manipulating either the forms or sample groups, which is done in the second study.

Research Questions

The following research questions are addressed in Study I:

(1) How do unidimensional IRT linking and equating methods perform for tests with various format effects?
(2) How do three unidimensional IRT linking methods compare?
(3) How do two unidimensional IRT equating methods compare?

Method

Data. Three Advanced Placement (AP) exams were selected based on the disattenuated correlation values between MC and FR scores. They were Comparative Government and Politics (Politics), English Language and Composition (English), and Spanish Literature (Spanish). In this study, data from the 2011 and 2012 administrations were used, and the 2011 form was the base (i.e., old) form and the 2012 form was the new form for each exam.

Table 1 shows descriptive statistics for the data used in this study. All three exams have both MC and FR items and there is a set of internal CIs consisting of MC items only. Hence, CIs are not format-representative. For the sake of simplicity, simple summed scores were used to construct composite scores, although non-integer weights for MC and FR sections are used with operational AP exams. Because the data were manipulated, this research does not pertain directly to AP exams, but is instead intended to address psychometric issues associated with mixed-format tests, in general.

To control for the effect of sample size, responses of 3,000 examinees were randomly sampled for each form. Table 2 provides the disattenuated correlation between MC and FR scores for the sampled data. The disattenuated correlation was .9s for Politics, .8s for English, and .7s for Spanish, indicating that format effects were smallest for Politics and largest for Spanish with English in the middle.

Table 2 also presents the CI effect size (ES), which is the standardized difference in the mean CI scores between the groups for the 2011 and 2012 forms. The CI ES was calculated using Equation 1 below (see Hagge & Kolen, 2012; Powers & Kolen, 2012):

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

where $\bar{x}_1$ is the mean for the new form, $\bar{x}_2$ is the mean for the old form, n_1 is the number of examinees for the new form, n_2 is the number of examinees for the old form, s_1^2 is the variance of the new form, and s_2^2 is the variance of the old form. As shown in Table 2, the CI ESs were quite small suggesting that the examinees who took the 2011 form were similar in their proficiency to those who took the 2012 form.

Analysis. First, the parameters of the two forms for each exam were estimated and linked by all three linking methods using flexMIRT 2.0 (Cai, 2013). For SEP, the Stocking-Lord method was conducted using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). The unidimensional three-parameter logistic model (Birnbaum, 1968) was used for the MC items, and the unidimensional graded response model (Samejima, 1969) was used for the FR items. In addition, prior distributions were used for the slope and pseudo-guessing parameters: lognormal (0, 0.5) for the slope parameters and beta ($\alpha = 5$, $\beta = 17$) for the pseudo-guessing parameters.

The IRT true score and observed score equating methods were performed to obtain raw-score equivalents, scale-score equivalents ranging from 0 to 70, and the AP grade equivalents of 1 to 5. The 0-70 scale scores for the old forms were created for this study using normalization. The AP grades for the old forms were also created for this study to have characteristics that are similar to those for operational grades. *Equating Recipes* (Brennan et al., 2009) was used for IRT true and observed score equating methods.

Evaluation. Equating results are evaluated for raw scores and scale scores. Also, examinee proportions of AP grade levels are compared. Because there is no criterion equating

relationship, results are compared to each other to see whether different methods lead to similar or dissimilar results. It is particularly interesting to compare results across three exams with different levels of format effects.

Results

Raw scores. Figures 1 through 3 show the raw-score equating relationships of three linking methods and two equating methods, for Politics, English, and Spanish, respectively. As shown in the figures, the absolute values of old-form equivalents minus new-form scores tend to be smallest for Politics and largest for Spanish overall, which means that the form difference was smallest for Politics and largest for Spanish.

For Politics in Figure 1, CON tended to perform somewhat differently from the other two linking methods although the differences were quite small. By contrast, for Spanish in Figure 3, the three linking methods showed quite different results in a wide score range between 10 and 80, and the differences were largest between SEP and FIX. This finding seems to indicate that the linking methods tend to perform more differently as the format effects get larger. However, there is a possibility that some confounding effects exist between the degree of form differences and format effects. Finally, for all exams, the two equating methods provided similar results, except near the very low raw-score points where true scores cannot be defined for raw scores lower than the sum of the pseudo-guessing parameter estimates.

Scale scores. Although not reported here, the unrounded scale-score equivalents were very similar across the linking and equating methods for all exams. However, the rounded scale-score equivalents showed some discrepancies. Tables 3 through 8 present the rounded scale-score equivalents based on the true and observed score equating methods for three exams. For Politics in Tables 3 and 4, there are 33 (37%, for true score) and 34 (38%, for observed score) out of 90 lines that show discrepancies across three linking methods. For English in Tables 5 and 6, 32 (39%, for true score) and 30 (36%, for observed score) out of 83 lines show discrepancies, and for Spanish, 48 (44%, for true score) and 56 (52%, for observed score) out of 108 lines. Similar to raw-score equating results, more discrepancies were observed for Spanish.

AP grades. Tables 9 to 11 show the proportion of examinees receiving each grade for Politics, English, and Spanish, respectively. Compared to the other two linking methods, SEP tends to show more differences across all exams, but the differences were not large (.0077 ~ .0280). Between the two equating methods, the results were almost the same.

Summary

According to all evaluation criteria (i.e., raw scores, scale scores, and AP grades), the two unidimensional IRT equating methods tended to perform in a similar way, and there did not appear to be an interaction between unidimensional IRT linking and equating methods. On the other hand, the three unidimensional IRT linking methods tended to show somewhat different results. For the raw-score and scale-score equating, CON performed differently than the other two linking methods when the format effects were small. When the format effects were large, all three linking methods showed somewhat different performance, and the differences between SEP and FIX were the largest. Based on the proportions of AP grades, SEP showed results that were different from the other two methods, regardless of the format effects; however, the differences were very small.

Study II: Pseudo Forms and Groups

An advantage of using the intact forms and groups in Study I is that the data are realistic. However, it is difficult to decide which method is associated with less equating error than the others because there is no absolute criterion. The second study based on pseudo forms and groups was conducted to evaluate the performance of the linking and equating methods relative to criterion equating relationships.

Research Questions

The research questions for Study II are as follows:

(1) How do the unidimensional IRT linking and equating methods perform relative to criterion equating relationships?

(2) How do the unidimensional IRT linking and equating methods perform under various testing conditions including format effects, form difficulty differences, CI compositions, CI proportions, and group differences?

Method

Construction of pseudo forms. The 2012 forms used in Study I were used to construct pseudo forms. To be specific, each of the three 2012 forms was split into two pseudo forms—a pseudo old form and a pseudo new form. This study was interested in understanding the effects of form difficulty differences, CI compositions, and CI proportions, as described previously. Therefore, these three factors were manipulated when creating pseudo forms. As a result, 14

distinct pairs of pseudo forms (2 form differences × 7 CI compositions and proportions) were constructed for each exam, and pseudo old forms were considered as the base forms.

Since the responses to the two pseudo forms were from the same examinees (the group that took the 2012 form), the single group equipercentile equating of scores on each pair of pseudo forms was considered as the criterion equating relationship. Table 12 presents the characteristics of the pseudo forms for the population.

Construction of pseudo groups. After creating the pseudo forms, pseudo groups were constructed to investigate the effect of group ability differences. In order to construct pseudo groups, first, two demographic variables related to test scores were selected. They were the educational level of examinees' fathers, which had 10 categories, and ethnicity with nine categories. Based on these variables, the population was divided into 90 categories (10 categories for the educational level × 9 categories for ethnicity). Therefore, each pseudo-old-form dataset and each pseudo-new-form dataset had 90 groups of examinees. However, because some of those 90 groups included a small number of examinees, several groups with similar CI ESs were combined to get larger datasets. Among those combined groups, three pairs of groups were chosen for each pair of pseudo forms. The groups were selected to produce three levels of CI ES (0.05, 0.25, and 0.50). Finally, in order to control for the effect of sample sizes, 3,000 examinees were randomly sampled to create each pseudo group.

Analysis. The analyses were similar to those used in Study I. The major difference was that scale scores between 0 and 25 were used for scale-score equating. This was because the test length was shorter in Study II by splitting a test form into two pseudo forms.

Factors of investigation. In addition to the three factors considered in Study I (i.e., linking methods, equating methods, and format effects), the second study investigates four additional factors: form differences, CI compositions, CI proportions, and group differences, which is summarized in Table 13.

Form differences were manipulated using the differences in *P*-values between the two pseudo forms. Two levels of form differences were considered: a small form difference with a *P*-value difference of .003, and a large form difference with a *P*-value difference of .04.

In Study I, CIs included MC items only, and CI proportions were about .2 ~ .3 of total scores. In order to examine the effects of CI compositions and CI proportions, seven levels were

investigated as follows: .1, .2, .3, and .4 for MC CIs, and .2, .3, and .4 for mixed CIs (MX CIs). Note that a CI proportion of .1 with MX CIs was not possible.

Finally, group differences were manipulated, which could not be investigated in Study I. Three levels of group differences were considered: CI ES of 0.05, 0.25, and 0.50.

Evaluation. The evaluation criteria were the single group equipercentile equating relationships. The criterion equating relationships were computed using the single group data. The following summary statistics were computed: standardized weighted averaged root mean squared difference (WARMSD), standardized unweighted averaged root mean squared difference (UARMSD), and average absolute difference in proportions of population examinees per grade level. Specifically, standardized WARMSD was computed for raw- and scale-score equivalents as:

$$Standardized\ WARMSD = \frac{1}{SD}\sqrt{\sum_i w_i[\hat{e}(x_i) - e^*(x_i)]^2}\,, \tag{2}$$

where SD is the standard deviation of the old form raw (or scale) scores, w_i is the proportion of examinees at the i-th new form raw score in the population, $\hat{e}(x_i)$ is the old form equivalent of the i-th new form raw score for a study condition equating relationship, and $e^*(x_i)$ is the old form equivalent for the criterion equating relationship.

Standardized UARMSD was computed for raw- and scale-score equivalents as:

$$Standardized\ UARMSD = \frac{1}{SD}\sqrt{\frac{\sum_i[\hat{e}(x_i) - e^*(x_i)]^2}{n}}\,, \tag{3}$$

where SD, $\hat{e}(x_i)$, and $e^*(x_i)$ are as defined previously, and n is the number of new form raw score points.

Last, the average absolute difference in proportions of population examinees per AP grade level was computed as $\left(\sum_{i=1}^{5}|p_i - p_i^*|\right)/5$, where p_i and p_i^* are the proportions of population examinees in grade i for a study condition equating relationship and for the criterion equating relationship, respectively.

Results

Raw scores. The aggregated (unweighted, over all the other conditions) standardized WARMSDs and UARMSDs for each study factor are shown from Tables 14 to Table 19. Note

that the results from two equating methods were not aggregated. Moreover, since the two equating methods showed the same pattern of results, only the observed score equating results are presented here.

First, in Table 14, the performance of three linking methods is generally similar (with CON performing slightly better) in terms of the standardized WARMSDs; however, the performance of FIX has noticeably more error than the other two methods in terms of the standardized UARMSDs. Second, Table 15 shows that Politics with the smallest format effects has the lowest error in terms of both statistics, while Spanish with the largest format effects is associated with the largest error. Third, as seen in Table 16, the large form-difference condition shows slightly larger standardized WARMSDs than the large form-difference condition; however, in terms of the standardized UARMSDs, the opposite is true. Thus, no definite conclusion could be made. Fourth, regarding CI compositions, Table 17 suggests that error is smaller for the MX CI condition than the MC CI condition. Fifth, error tends to decrease as the CI proportion increases as seen in Table 18. Sixth, as the group difference increases, error tends to increase as well, as shown in Table 19.

Table 20 presents two-way interaction effects between the linking methods and format effects. The format effects seem to have larger influence on SEP and FIX than on CON. In particular, the comparison between SEP and CON shows that, based on both WARMSD and UARMSD statistics, SEP performs slightly better than CON for Politics and English. However, for Spanish, CON tends to perform better than SEP. Compared to FIX, CON provides lower error for almost all cases. This finding is surprising because it has been speculated that CON may be sensitive to multidimensionality (Kolen & Brennan, 2014). Table 21 presents three-way interaction effects between the linking methods, format effects, and CI Compositions. It appears that when MX CIs are used, SEP and FIX provide much lower error for Spanish than when MC CIs are used. This indicates that use of mixed-format CIs would be more beneficial for SEP and FIX than CON.

Thus far, results have been provided in an aggregated manner. In order to examine the results across the score scale, conditional differences between the criterion and study conditions could be plotted. Due to the space limit, only one set of study conditions, which was judged to be most typical, was included. Figures 4 to 6 are plots for the conditional differences in equivalents between a set of study conditions and criterion equating relationships for Politics, English, and

Spanish, respectively. The selected study condition was: observed score equating, small form difference, MC 20% CIs, and CI ES of 0.05. The two dotted lines on the plots represent a band for the standardized differences that matter (DTM; Dorans & Feigenbaum, 1994). Note that there is a range of raw scores where error for CON goes beyond the DTM lines (Figure 6).

Scale scores. Scale-score equating results were almost identical to the raw-score equating results. Thus, they are not discussed here.

AP grades. AP grade proportion results also were similar to the raw-score equating results in many cases. However, the effects of format effects and form differences were different and discussed here. Table 22 shows the results for format effects. The largest error is associated with Spanish, and the smallest with English. Note that the difference between Politics and English is only 0.0011. Second, Table 23 presents the effect of form differences. The large form-difference condition clearly shows larger error than the small form-difference condition.

Summary

In general, two unidimensional IRT equating methods showed similar performance in terms of the robustness to format effects, which is consistent with the results reported in Study I. Lower error was associated with the following conditions: mixed CIs, high CI proportions, small group difference, and large form difference (partially supported). Three unidimensional IRT linking methods performed somewhat differently. In particular, compared to CON, SEP and FIX were more sensitive to format effects resulting in larger error for Spanish. However, when both MC and FR items were used as CIs, SEP and FIX showed much better performance even when format effects were large. Politics with the smallest format effects has the lowest error in terms of both statistics, while Spanish with the largest format effects is associated with the largest error.

Conclusions and Discussion

Two studies were conducted in this paper. In the first study using data from intact forms and groups, equating results were compared for three linking methods, two equating methods, and three exams with various degrees of format effects. The three linking methods showed larger differences in equating results for the exams with large format effects (i.e., English and Spanish) than for the exam with small format effects (i.e., Politics). This finding suggests that the degree of format effects is one of the factors that should be considered when choosing among various linking methods for mixed-format tests.

The second study used pseudo forms and groups, and considered seven study factors: linking methods, equating methods, format effects, form differences, CI compositions, CI proportions, and group differences. Equating results were obtained for raw scores, scale scores, and AP grades. Some of the major findings for the six study factors were either consistent with previous studies or as expected, in general. The two equating methods showed similar performance. Larger error was associated with larger format effects, lower CI proportions, and larger group differences. Moreover, the mixed-CI conditions almost always produced better results than the MC only CI conditions across other conditions.

The findings of the second study suggest that CON tends to be less vulnerable to equating error associated with format effects. Specifically, SEP tended to perform better than CON for exams with smaller format effects (i.e., Politics and English); however, CON outperformed SEP for Spanish, which has the largest format effects. This finding is somewhat inconsistent with the results from previous studies. For example, when data do not fit a unidimensional IRT model (e.g., due to multidimensionality), SEP was associated with less error than CON (Béguin & Hanson, 2001). Although CON was associated with less error than SEP and FIX with large format effects, the relative performance of SEP and FIX tended to improve noticeably when mixed CIs were used. This finding suggests that, when a large format effects exist, the use of mixed CIs will be much more effective than the use of MC CIs for SEP and FIX.

Some limitations should be recognized, especially for the second study. First, the second study used the single group equipercentile equating relationships as criteria. It is possible that using different criteria would lead to different conclusions. Future research could employ simulation using multidimensional models and criterion equating relationships could be determined based on the generating models. Second, only one set of data was created for the pseudo forms and groups. A future study would involve replicating the sampling process a number of times to obtain more stable results. Future studies could also be conducted using pseudo forms of other exams.

References

Andrews, B. J. (2011). *Assessing first- and second-order equity for the common-item nonequivalent groups design using multidimensional IRT* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

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 Number 1). Iowa City, IA: CASMA, The University of Iowa.

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

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

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. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Hagge, S. L., & Kolen, M. J. (2012). Effects of group differences on equating using operational and pseudo-tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (CASMA Monograph Number. 2.2) (pp. 45-86). Iowa City, IA: CASMA, The University of Iowa.

He, Y., & Kolen, M. J. (2011). Equity and same distributions properties for test 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. 177-212). Iowa City, IA: CASMA, The University of Iowa.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

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

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*(2), 15-24.

Li, Y. H., Tam, H. P., & Tompkins, L. J. (2004). A comparison of using the fixed common-precalibrated parameter method and the matched characteristic curve method for linking multiple-test items. *International Journal of Testing, 4*, 267-293.

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

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

Powers, S., & Kolen, M. J. (2012). Using matched samples equating methods to improve equating accuracy. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (CASMA Monograph Number 2.2) (pp. 87-114). Iowa City, IA: CASMA, The University of Iowa.

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

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

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Table 1

Descriptive Information for Exams Used in Study I

Exam	Number of MC Items	Number of FR Items	Maximum FR Score	Total Raw Composite Score	Number of MC CIs
Politics 2011	55	8	3, 3, 3, 3, 3, 5, 7, 8	90	18
Politics 2012	55	8	3, 3, 3, 3, 3, 5, 7, 7	89	
English 2011	54	3	9 each	81	25
English 2012	55	3	9 each	82	
Spanish 2011	65	6	9, 5, 9, 5, 9, 5	107	22
Spanish 2012	64	6	9, 5, 9, 5, 10, 5	107	

Table 2

Disattenuated Correlation Between MC and FR Scores and CI Effect Sizes

Exam	Disattenuated Correlation	CI Effect Size
Politics 2011	.9499	0.0789
Politics 2012	.9672	
English 2011	.8153	-0.0267
English 2012	.8234	
Spanish 2011	.7606	0.0094
Spanish 2012	.7731	

Table 3

Rounded Scale Score Equivalents of IRT True Score Equating for Politics

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	2	2	2	30	23	23	23	60	41	41	41
1	3	3	3	31	24	23	24	61	41	42	41
2	4	4	4	32	24	24	24	62	42	42	42
3	4	4	4	33	25	25	25	63	43	43	43
4	5	5	5	34	26	25	26	64	43	44	43
5	5	5	5	35	26	26	26	65	44	45	44
6	6	6	6	36	27	27	27	66	45	45	45
7	7	7	7	37	27	27	27	67	46	46	46
8	7	7	7	38	28	28	28	68	47	48	47
9	8	8	8	39	28	28	28	69	48	49	48
10	8	9	8	40	29	29	29	70	49	50	49
11	9	9	9	41	29	29	29	71	50	51	50
12	10	10	10	42	30	30	30	72	51	52	51
13	10	10	10	43	30	30	30	73	52	53	52
14	11	11	11	44	31	31	31	74	53	54	53
15	12	12	12	45	31	31	31	75	54	55	54
16	13	13	13	46	32	32	32	76	55	56	55
17	14	13	13	47	32	32	32	77	56	57	56
18	14	14	14	48	33	33	33	78	57	58	57
19	15	15	15	49	33	34	34	79	58	59	58
20	16	16	16	50	34	34	34	80	59	60	59
21	17	16	16	51	35	35	35	81	60	61	60
22	17	17	17	52	35	36	36	82	62	62	61
23	18	18	18	53	36	36	36	83	63	63	62
24	19	19	19	54	37	37	37	84	64	64	63
25	20	19	19	55	37	38	38	85	65	65	64
26	20	20	20	56	38	38	38	86	66	66	66
27	21	21	21	57	39	39	39	87	67	67	67
28	22	21	22	58	39	40	39	88	69	69	68
29	22	22	22	59	40	40	40	89	70	70	70

Note. The shaded cells indicate different equivalents across linking methods.

Table 4

Rounded Scale Score Equivalents of IRT Observed Score Equating for Politics

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	3	3	3	30	23	23	23	60	41	41	41
1	3	3	3	31	24	23	24	61	41	42	41
2	3	3	3	32	24	24	24	62	42	42	42
3	4	4	4	33	25	25	25	63	43	43	43
4	5	5	5	34	26	25	26	64	43	44	43
5	5	5	5	35	26	26	26	65	44	45	44
6	6	6	6	36	27	27	27	66	45	45	45
7	7	7	7	37	27	27	27	67	46	46	46
8	7	7	7	38	28	28	28	68	47	47	47
9	8	8	8	39	28	28	28	69	48	48	48
10	9	9	9	40	29	29	29	70	49	50	49
11	10	9	9	41	29	29	29	71	50	51	50
12	10	10	10	42	30	30	30	72	51	52	51
13	11	11	11	43	30	30	30	73	52	53	52
14	12	11	11	44	31	31	31	74	53	54	53
15	12	12	12	45	31	31	31	75	54	55	54
16	13	13	13	46	32	32	32	76	55	56	55
17	14	14	14	47	32	32	32	77	56	57	56
18	15	14	14	48	33	33	33	78	57	58	57
19	15	15	15	49	33	34	34	79	58	59	58
20	16	16	16	50	34	34	34	80	59	60	59
21	17	16	17	51	35	35	35	81	60	61	60
22	18	17	17	52	35	36	36	82	61	62	61
23	18	18	18	53	36	36	36	83	62	63	62
24	19	19	19	54	37	37	37	84	63	64	63
25	20	19	19	55	37	38	38	85	65	65	64
26	20	20	20	56	38	38	38	86	66	66	65
27	21	21	21	57	39	39	39	87	67	67	66
28	22	21	22	58	39	40	40	88	68	68	68
29	22	22	22	59	40	40	40	89	69	69	69

Note. The shaded cells indicate different equivalents across linking methods.

Table 5

Rounded Scale Score Equivalents of IRT True Score Equating for English

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	0	0	0	28	19	19	19	56	44	44	44
1	1	1	1	29	20	20	20	57	45	45	45
2	1	1	1	30	21	21	21	58	46	46	46
3	2	2	2	31	22	22	22	59	47	47	47
4	3	3	3	32	23	23	23	60	47	47	47
5	4	4	3	33	24	24	23	61	48	48	48
6	4	4	4	34	25	24	24	62	49	49	49
7	5	5	5	35	25	25	25	63	50	50	50
8	6	6	5	36	26	26	26	64	52	51	51
9	7	6	6	37	27	27	27	65	53	53	52
10	7	7	6	38	28	28	28	66	54	54	53
11	8	7	7	39	29	29	29	67	55	55	54
12	8	8	7	40	30	30	30	68	56	56	55
13	9	8	8	41	31	31	31	69	57	56	56
14	9	9	8	42	32	32	32	70	57	57	57
15	10	10	9	43	33	32	32	71	58	58	58
16	11	10	10	44	34	33	33	72	59	59	59
17	11	11	10	45	35	34	34	73	60	60	60
18	12	12	11	46	36	35	35	74	61	61	61
19	13	12	12	47	37	36	36	75	62	62	62
20	13	13	13	48	38	37	37	76	63	63	63
21	14	14	13	49	38	38	38	77	64	64	64
22	15	15	14	50	39	39	39	78	65	65	65
23	16	15	15	51	40	40	40	79	66	66	66
24	16	16	16	52	41	41	41	80	67	67	67
25	17	17	17	53	42	42	42	81	68	68	68
26	18	18	17	54	43	43	43	82	70	70	70
27	19	18	18	55	44	44	44				

Note. The shaded cells indicate different equivalents across linking methods.

Table 6

Rounded Scale Score Equivalents of IRT Observed Score Equating for English

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	0	0	0	28	20	20	19	56	44	44	44
1	0	0	0	29	21	20	20	57	45	45	45
2	0	0	0	30	21	21	21	58	46	46	46
3	2	1	0	31	22	22	22	59	47	47	47
4	2	2	2	32	23	23	23	60	47	47	47
5	3	3	2	33	24	24	24	61	48	48	48
6	4	3	3	34	25	25	25	62	49	49	49
7	4	4	4	35	26	25	25	63	51	50	50
8	5	5	4	36	27	26	26	64	52	52	51
9	6	5	5	37	27	27	27	65	53	53	52
10	6	6	6	38	28	28	28	66	54	54	53
11	7	7	6	39	29	29	29	67	55	55	54
12	8	7	7	40	30	30	30	68	56	56	56
13	8	8	8	41	31	31	31	69	57	57	57
14	9	9	8	42	32	32	32	70	58	58	58
15	10	10	9	43	33	32	32	71	59	59	59
16	11	10	10	44	34	33	33	72	60	60	60
17	11	11	11	45	35	34	34	73	61	61	61
18	12	12	11	46	36	35	35	74	62	62	62
19	13	12	12	47	36	36	36	75	63	63	63
20	13	13	13	48	37	37	37	76	64	64	64
21	14	14	14	49	38	38	38	77	65	65	64
22	15	15	14	50	39	39	39	78	66	66	66
23	16	16	15	51	40	40	40	79	67	67	67
24	17	16	16	52	41	41	41	80	68	68	68
25	17	17	17	53	42	42	42	81	69	69	69
26	18	18	18	54	43	43	43	82	70	70	70
27	19	19	19	55	44	44	43				

Note. The shaded cells indicate different equivalents across linking methods.

Table 7

Rounded Scale Score Equivalents of IRT True Score Equating for Spanish

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	4	4	4	36	17	17	16	72	32	32	32
1	4	4	4	37	18	17	17	73	33	33	33
2	5	5	5	38	18	17	17	74	34	34	34
3	5	5	5	39	18	18	17	75	35	35	35
4	5	5	5	40	19	18	18	76	36	36	36
5	6	6	6	41	19	19	18	77	37	36	36
6	6	6	6	42	19	19	19	78	37	37	37
7	7	7	6	43	20	19	19	79	38	38	38
8	7	7	7	44	20	20	19	80	39	39	39
9	7	7	7	45	21	20	20	81	40	40	40
10	8	8	7	46	21	20	20	82	41	41	41
11	8	8	8	47	21	21	21	83	42	42	42
12	9	9	8	48	22	21	21	84	43	43	43
13	9	9	9	49	22	22	21	85	44	44	44
14	9	9	9	50	22	22	22	86	45	45	45
15	10	10	9	51	23	22	22	87	46	46	46
16	10	10	9	52	23	23	23	88	47	47	47
17	10	10	10	53	23	23	23	89	48	48	48
18	11	11	10	54	24	23	23	90	49	50	50
19	11	11	10	55	24	24	24	91	51	51	51
20	12	11	11	56	25	24	24	92	52	52	52
21	12	11	11	57	25	25	25	93	53	53	53
22	12	12	11	58	25	25	25	94	54	54	54
23	13	12	12	59	26	25	25	95	55	56	56
24	13	12	12	60	26	26	26	96	57	57	57
25	13	13	12	61	26	26	26	97	58	58	58
26	14	13	13	62	27	27	27	98	59	59	59
27	14	13	13	63	27	27	27	99	60	61	61
28	14	14	13	64	28	27	27	100	61	62	62
29	15	14	14	65	28	28	28	101	63	63	63
30	15	15	14	66	29	28	28	102	64	64	64
31	15	15	14	67	29	29	29	103	65	65	65
32	16	15	15	68	30	30	30	104	66	66	66
33	16	16	15	69	30	30	30	105	67	68	68
34	17	16	16	70	31	31	31	106	69	69	69
35	17	16	16	71	32	31	31	107	70	70	70

Note. The shaded cells indicate different equivalents across linking methods.

Table 8

Rounded Scale Score Equivalents of IRT Observed Score Equating for Spanish

Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX	Raw Score	SEP	CON	FIX
0	5	5	4	36	17	17	16	72	32	32	32
1	5	5	4	37	18	17	17	73	33	33	33
2	5	5	4	38	18	17	17	74	34	34	34
3	5	5	4	39	18	18	17	75	35	35	35
4	5	5	4	40	19	18	18	76	36	36	36
5	5	5	4	41	19	18	18	77	37	36	36
6	6	6	4	42	19	19	19	78	38	37	37
7	7	6	6	43	20	19	19	79	38	38	38
8	7	7	6	44	20	20	19	80	39	39	39
9	7	7	7	45	20	20	20	81	40	40	40
10	8	7	7	46	21	20	20	82	41	41	41
11	8	8	7	47	21	21	21	83	42	42	42
12	8	8	8	48	22	21	21	84	43	43	43
13	9	8	8	49	22	21	21	85	44	44	44
14	9	9	8	50	22	22	22	86	45	45	45
15	10	9	9	51	23	22	22	87	46	46	46
16	10	10	9	52	23	23	23	88	47	47	47
17	10	10	9	53	23	23	23	89	48	48	48
18	11	10	10	54	24	23	23	90	49	49	49
19	11	11	10	55	24	24	24	91	50	50	50
20	11	11	10	56	24	24	24	92	52	52	52
21	12	11	11	57	25	25	25	93	53	53	53
22	12	12	11	58	25	25	25	94	54	54	54
23	12	12	11	59	26	25	25	95	55	55	55
24	13	12	12	60	26	26	26	96	56	56	56
25	13	13	12	61	26	26	26	97	57	58	58
26	14	13	13	62	27	26	26	98	59	59	59
27	14	13	13	63	27	27	27	99	60	60	60
28	14	14	13	64	28	27	27	100	61	61	61
29	15	14	14	65	28	28	28	101	62	62	62
30	15	14	14	66	29	28	28	102	63	63	64
31	15	15	14	67	29	29	29	103	64	65	65
32	16	15	15	68	30	30	30	104	66	66	66
33	16	16	15	69	30	30	30	105	67	67	67
34	16	16	16	70	31	31	31	106	68	68	68
35	17	16	16	71	32	31	31	107	69	69	69

Note. The shaded cells indicate different equivalents across linking methods.

Table 9

AP Grade Distributions for Politics

AP Grade	Old Form Distribution	SEP		CON		FIX	
		True Score	Observe Score	True Score	Observe Score	True Score	Observe Score
1	.1873	.1693	.1693	.1693	.1693	.1693	.1693
2	.2527	.2450	.2450	.2223	.2223	.2223	.2223
3	.1760	.1747	.1747	.1973	.1973	.1973	.1973
4	.2453	.2513	.2513	.2513	.2513	.2513	.2513
5	.1387	.1597	.1597	.1597	.1597	.1597	.1597

Table 10

AP Grade Distributions for English

AP Grade	Old Form Distribution	SEP		CON		FIX	
		True Score	Observe Score	True Score	Observe Score	True Score	Observe Score
1	.0997	.1380	.1173	.1380	.1380	.1380	.1380
2	.2367	.2093	.2300	.2093	.2093	.2093	.2093
3	.2640	.2437	.2437	.2437	.2437	.2437	.2437
4	.2560	.2450	.2623	.2623	.2623	.2623	.2623
5	.1437	.1640	.1467	.1467	.1467	.1467	.1467

Table 11

AP Grade Distributions for Spanish

AP Grade	Old Form Distribution	SEP		CON		FIX	
		True Score	Observe Score	True Score	Observe Score	True Score	Observe Score
1	.3207	.3207	.3207	.3410	.3410	.3410	.3410
2	.2293	.2240	.2240	.2317	.2317	.2317	.2317
3	.2400	.2543	.2543	.2263	.2263	.2263	.2263
4	.1557	.1473	.1473	.1473	.1473	.1473	.1473
5	.0543	.0537	.0537	.0537	.0537	.0537	.0537

Table 12

Pseudo Form Information for Populations

Exam	Number of Examinees	Summed Raw Score	Maximum MC Score	Maximum FR Score
Politics	17,750	40	24	16
English	20,000	30	21	9
Spanish	17,527	50	30	20

Table 13

Factors of Investigation for Study II

Factor	Description
Calibration methods	Separate calibration with Stocking and Lord method (SEP) Concurrent calibration (CON) Fixed parameter calibration (FIX)
Equating methods	IRT true score equating IRT observed score equating
Format effects/Exams	Politics English Spanish
Form differences	Small (difference in *P*-values of .0030) Large (difference in *P*-values of .0400)
CI compositions	MC items only (MC CIs) MC and FR items (MX CIs)
CI proportions	.1, .2, .3, and .4 for MC CIs .2, .3, and .4 for MX CIs
Group differences	CI ES of 0.05, 0.25, and 0.50

Table 14

Raw-Score Equating Results for Linking Methods

	SEP	CON	FIX
WARMSD	0.0465	0.0429	0.0468
UARMSD	0.0567	0.0541	0.0750

Table 15

Raw-Score Equating Results for Format Effects

	Politics	English	Spanish
WARMSD	0.0320	0.0398	0.0644
UARMSD	0.0421	0.0490	0.0947

Table 16

Raw-Score Equating Results for Form Differences

	Small	Large
WARMSD	0.0448	0.0460
UARMSD	0.0632	0.0606

Table 17

Raw-Score Equating Results for CI Compositions

	MC	MX
WARMSD	0.0456	0.0347
UARMSD	0.0637	0.0512

Table 18

Raw-Score Equating Results for CI Proportions

	20%	30%	40%
WARMSD	0.0461	0.0398	0.0345
UARMSD	0.0609	0.0587	0.0527

Table 19

Raw-Score Equating Results for Group Differences

	ES 0.05	ES 0.25	ES 0.50
WARMSD	0.0346	0.0419	0.0597
UARMSD	0.0496	0.0579	0.0782

Table 20

Raw-Score Equating Results for Linking Methods and Format Effects/Exams

	Exam	Linking Methods		
		SEP	CON	FIX
WARMSD	Politics	0.0289	0.0336	0.0335
	English	0.0354	0.0416	0.0425
	Spanish	0.0752	0.0536	0.0643
UARMSD	Politics	0.0355	0.0425	0.0481
	English	0.0445	0.0484	0.0540
	Spanish	0.0900	0.0713	0.1227

Table 21

Raw-Score Equating Results for Linking Methods, Format Effects/Exams, and CI Compositions

	Exam	CI Composition	Linking Method		
			SEP	CON	FIX
WARMSD	Politics	MC	0.0278	0.0302	0.0299
		MX	0.0214	0.0250	0.0300
	English	MC	0.0324	0.0343	0.0382
		MX	0.0312	0.0354	0.0361
	Spanish	MC	0.0925	0.0498	0.0755
		MX	0.0368	0.0437	0.0528
UARMSD	Politics	MC	0.0335	0.0376	0.0389
		MX	0.0318	0.0385	0.0511
	English	MC	0.0467	0.0453	0.0525
		MX	0.0380	0.0408	0.0430
	Spanish	MC	0.0990	0.0723	0.1473
		MX	0.0565	0.0636	0.0977

Table 22

Grade Proportion Results for Format Effects/Exams

Politics	English	Spanish
0.0082	0.0071	0.0122

Table 23

Grade Proportion Results for Form Differences

Small	Large
0.0055	0.0128

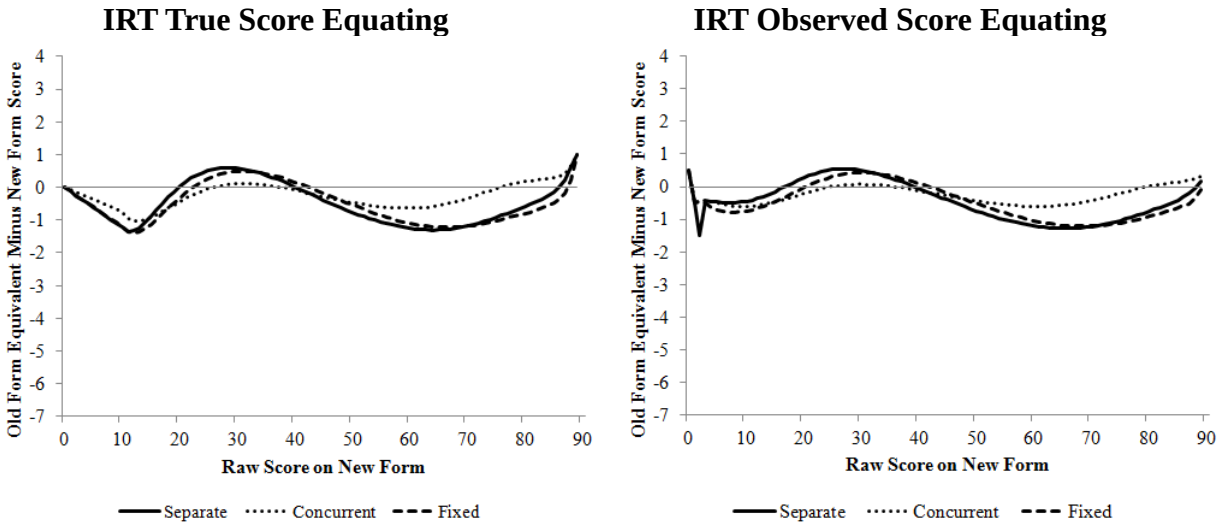

Figure 1. Raw-score equating relationships for Politics.

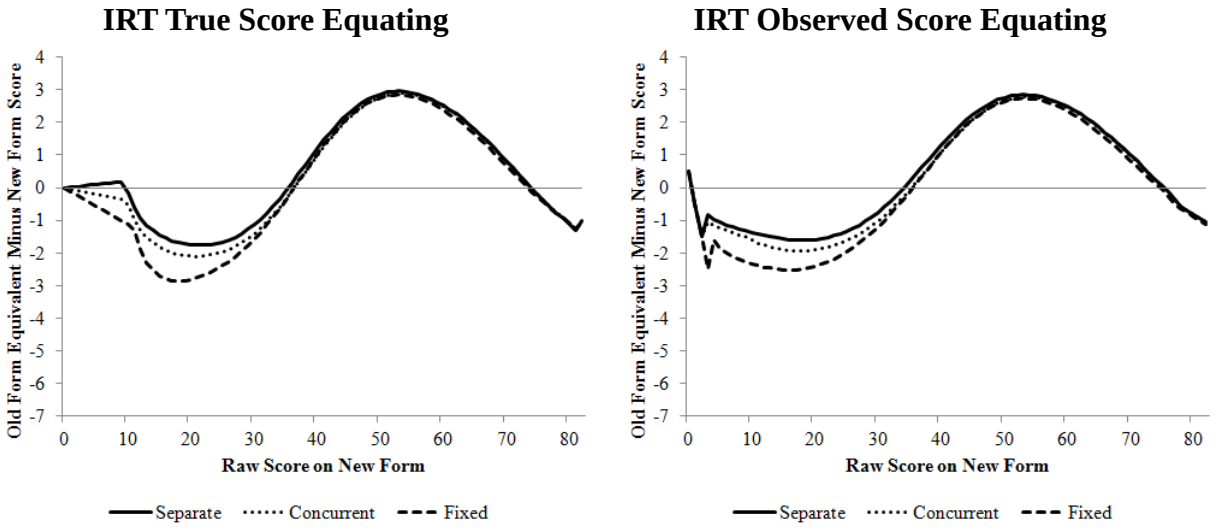

Figure 2. Raw-score equating relationships for English.

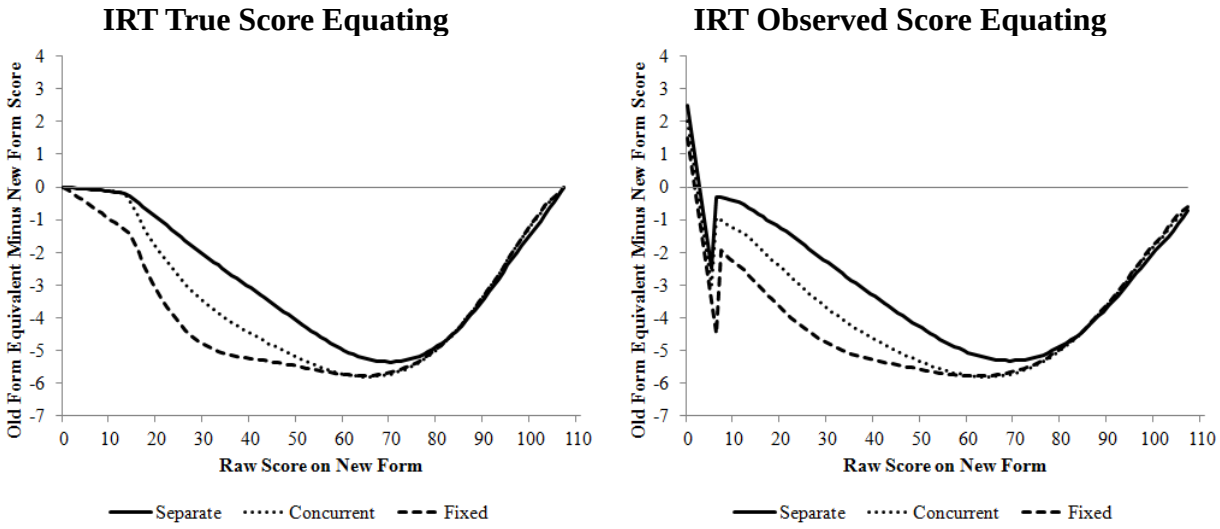

Figure 3. Raw-score equating relationships for Spanish.

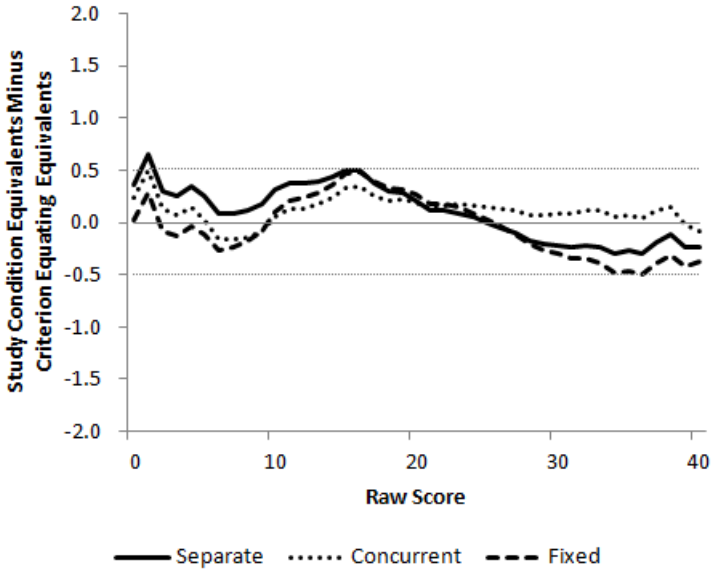

Figure 4. Differences between the study condition and criterion equating relationships for Politics.

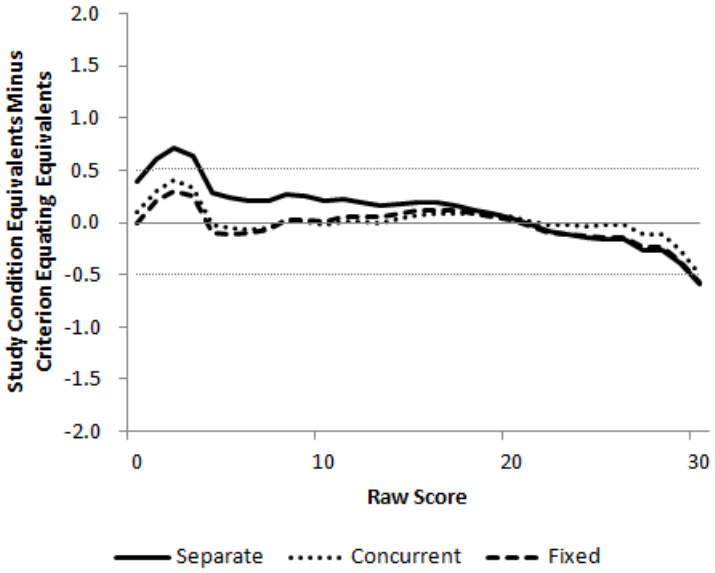

Figure 5. Differences between the study condition and criterion equating relationships for English.

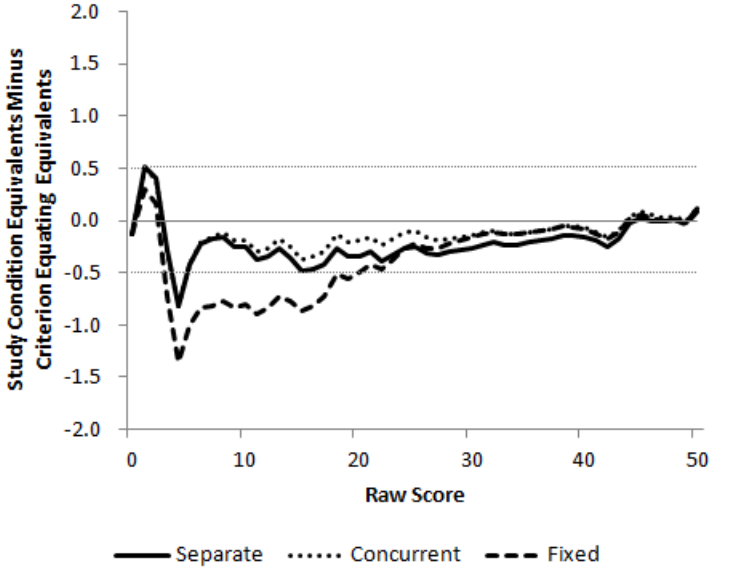

Figure 6. Differences between the study condition and criterion equating relationships for Spanish.

Chapter 4: Evaluating the Interchangeability of Free-Response Items Developed from Task Models

Jaime L. Malatesta and Huan Liu

The University of Iowa, Iowa City, IA

Abstract

The purpose of this study was to evaluate the interchangeability of free-response (FR) items built from the same task model. Methods originally designed to detect unstable multiple-choice common items as well as a differential item functioning detection method were adapted and applied to the FR items. The specific methods used included: squared differences of item characteristic curves, the robust z method, ordinal logistic regression, and visual inspection of IRT item parameter estimates. Three Advanced Placement world language tests from four administrations were used: German, French, and Italian. Each test is mixed-format and contains four FR items, each of which, were developed from distinct task models. In general, the flagging methods produced consistent results and typically suggested that at least one of the four task models produced items that could be regarded as interchangeable from a psychometric standpoint.

Evaluating the Interchangeability of Free-Response Items Developed from Task Models

In evidence-centered design (ECD), task models play a central role in item development and provide a direct link between items and specific claims we wish to make about student performance and the evidence elicited during the test taking process. It follows that the use of task models can help support theory-driven test construction and item development by explicitly mapping items to measurement targets. From a test development perspective, the benefits of implementing ECD (together with task models) include the improvement in construct equivalence across different forms and the ability to generate a plethora of items from a single task model without much additional resources. From a teacher's perspective, the benefits include being able to use tests that are more closely aligned to what is taught and valued in the classroom as well as being able to link student test performance to curriculum-based learning objectives. Furthermore, from a validity perspective, score inferences are strongly supported by evidentiary arguments under the ECD approach (Hendrickson, Ewing, Kaliski, & Huff, 2013).

As previously mentioned, a single task model can be used to develop multiple items. The collection of items developed from the same task model is considered an item family where all items measure the same latent construct and have nearly identical psychometric properties (i.e., difficulty and discrimination). Traditionally, item development is a detailed process that requires subject matter experts to draft items, which then go through a thorough and iterative review process prior to being field tested. After field testing, items are reviewed again for fairness and statistical adequacy and depending on the outcome, are either approved for operational use, go through more revisions and additional field testing, or in some cases, are retired or released. The additional preparation work that goes into item development in ECD is intended to help increase the number of items that have good field test statistics and hence are ready for operational use. If task models are implemented with fidelity, some argue that the entire practice of field testing may no longer be needed (Luecht, 2013).

Under the ECD framework, item development is often even more resource-intensive and involves each of the aforementioned steps in addition to other steps. For example, additional time is spent integrating learning theory into the assessment design, identifying appropriate levels of specificity for the claims and evidence, and creating task models that can generate a desired number of items. While these additional requirements make ECD more time consuming to

implement initially, the benefits associated with being able to produce a collection of items with construct equivalence are appealing. Also appealing is the added utility of being able to populate item pools with quasi-parallel items, thereby enhancing the comparability of future test forms (Hendrickson, Huff, & Luecht, 2010). Under the most ideal setting, task models can lead to items with such similar characteristics that they might be used as common items (CIs) for linking and equating purposes. However, these uses hinge on the assumption that task models actually produce items that can be regarded as interchangeable from a content and psychometric standpoint.

Luecht (2009) outlined several features that task models must possess in order to be considered complete. First, they must describe performance expectations and conditions under which the task will be performed. Second, they need to be developed such that they occupy a particular point on an ability scale. Third, they need to contain features and data objects that fit into a template which can be systematically modified and used to generate a family of items. Therefore, a task model should give rise to multiple templates, and each template and the items produced by it, should function so similarly from a psychometric standpoint that they may be treated as isomorphs or identical items. It should be noted that the difference between a task model and a task template is the level of specificity at which they are written, with the latter being more specific. For the purpose of this study, it is assumed that task models are specific enough to give rise to item families.

Since each task model is associated with particular point on a measurement scale, all task templates and subsequent items derived from it should have similar psychometric properties (Luecht, 2009). Therefore, items belonging to the same task model should have psychometric properties that are more similar to one another than with items developed from different task models. At the very least, this is a relationship that one would expect to find if indeed, task models and their subsequent items were functioning well. If any of the aforementioned benefits, such as improved construct equivalence of test forms, enhanced validity evidence, or rapid item generation are to be realized, then evaluation criteria need to be developed in order to determine whether task models and their items are functioning as hypothesized. As Hendrickson et al. (2013) point out, there are no criteria available to evaluate task models, and whether the items generated from them function well. This study aims to address this issue by demonstrating how

several well-known item analysis procedures can be adapted to evaluate the interchangeability of free-response items built from the same task model.

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 (French and German) or 70 (Italian) multiple-choice (MC) items, along with 4 free-response (FR) items. The MC items are scores dichotomously and the FR items are each scored on a 6 point scale (i.e., 0, 1, 2, 3, 4, 5). The first FR item is an interpersonal writing task, the second is a presentational writing task, the third is actually a culmination of five short interpersonal speaking tasks aggregated into one score, and the fourth FR item represents a presentational speaking task. The purpose of this study is solely for research and therefore operational weights were not used and results should not be generalized to the operational AP exams. Instead, the maximum number of raw-score points for the French, German, and Italian exams were 85, 85, and 90, respectively.

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 back to 2012, 2014 linked back to 2012, and 2015 linked back to 2013.

Original exam data were collected under the common item nonequivalent groups (CINEG) design and as a result, examinee and form differences were confounded. In order to focus on form differences, pseudo-groups were created to approximate a random groups (RG) design. Therefore, two different datasets were used in this study (CINEG and RG), which effectively doubled the number of studied conditions to 18 links (3 subjects, 3 links per subject, replicated across two data collection designs).

In order to create RG data, a matching variable was created that represented the highest level of education that either the examinee's mother or father had achieved (coded 1-9; see Table 1). Stratified random sampling without replacement was conducted on both the original (i.e., CINEG data) old and new form samples in an iterative manner until the effect size for group differences, calculated using common item total scores, was less than |0.01|.

Because common item scores were used during the formation of the RG data, multiple samples had to be drawn from the 2012 and 2013 forms in order to mimic the original linking design. For example, the 2013 and 2014 new forms each link back to the 2012 old form, but the

items common between the 2012 and 2013 forms are different than those common between the 2012 and 2014 forms. Therefore, the 2012 form had to be sampled twice in order to create randomly equivalent groups for the 2012-2013 and 2012-2014 links. Using common-item scores as the dependent measure, it was impossible to sample a pseudo-group from the 2012 old form that would be approximately equal in ability to both the 2013 and 2014 pseudo-groups, simultaneously. As a result, two pseudo-group samples were drawn from the 2012 old form. Similarly, two pseudo-groups were sampled from the 2013 form; one served as the new form in the 2012-2013 link and the second served as the old form in the 2013-2015 link.

IRT Item Calibration and Scale Linking

All forms were calibrated using flexMIRT 3.0 (Cai, 2015). The MC and FR items were calibrated using the three-parameter logistic (3PL; Birnbaum, 1968) and Muraki's (1992) generalized partial credit models, respectively. 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 Stocking and Lord (1983) linking coefficients. The linking coefficients were then applied to the new form item parameter estimates in order to place them on the old form scale. It should be noted that scale transformation was only performed on the CINEG datasets and not the RG datasets. Also worth pointing out is that the 2015 form was linked back to the 2013 form, which was previously linked back to the 2012 form, and therefore represents a chain of linkings.

Evaluation Criteria

Once group differences were accounted for, by performing a scale transformation with the CINEG data and by using pseudo-groups to approximate the RG design, several indices were computed to evaluate whether task-model-derived FR items behaved similarly across forms. In order to do this, FR items that belonged to the same task model but that appeared on different forms, were treated as common items. For example, on each subject test, the first FR item on each form is built from the same task model and therefore, FR 1 on the 2012 form was treated as though it reappeared as FR 1 on the 2013 form. However, it is very important to point out that only operational MC common items were used to link the new forms to the old form scales and

that FR items were treated as common only after the linking was completed. These decisions were made largely because (a) there was no compelling evidence to prematurely treat any of the task-model-derived FR items as common items during the linking process and (b) if items developed from the same task model are to be treated as interchangeable for linking and equating purposes then they ought to be evaluated with the same degree of scrutiny as is done with actual common items.

Squared differences. A weighted squared difference (i.e., d^2) between item characteristic curves based on old (i.e., Y) and new form (i.e., X) parameter estimates was first computed as:

$$d_i^2 = \sum^{K} [P_{ix}(\theta_k) - P_{iy}(\theta_k)]^2 * g(\theta_k). \quad (1)$$

Here, $P_{ix}(\theta_k)$ represents the response probability for item *i* on Form X for an examinee with ability (θ) level associated with the k^{th} quadrature point, P_{iy} is a similar quantity but for Form Y, and $g(\theta_k)$ is the discrete density associated with the k^{th} quadrature point. This weighted squared difference is summed over $k = 1$ to K quadrature points. In this study, a theoretical normal (0, 1) posterior theta distribution with 101 quadrature points, spanning -5 to +5 was used.

The d^2 statistic was first computed for each operational common item. The mean and standard deviation of these d^2 values was then computed and used for flagging any FR items that had a d^2 value greater than two standard deviations above the mean. Since all operational common items were MC format, the scoring function of the FR items had to be placed on the same 0/1 scale. This was accomplished by dividing each cumulative FR response probability (which ranged from 0 to 5) by the number of response options (i.e., 5). Another way to work around this issue could have been to forgo any rescaling and to instead, compute squared differences on the category response probabilities (which range from 0 to 1 like MC items). Each FR item would have five d^2 values, corresponding to scores of 1, 2, 3, 4, or 5, which could then each be compared against the flagging criterion. Depending on one's desired level of statistical power, an item could be flagged if one, two, three, four, or all five of the d^2 values exceeded the two standard deviations criterion.

The d^2 statistic is commonly used in operational settings to check the stability of common items (Pearson, 2012) and has the benefits of requiring little computation and allowing for straightforward interpretation. However, it is model-based and therefore its usefulness hinges on the assumption that the IRT models adequately fit the data.

Robust z. The Robust z method was first introduced by Huynh (2000) and has been used by several large-scale state testing programs such as South Carolina, Arkansas, Maryland, Minnesota, and New Mexico to check for unstable common items in conjunction with linking and equating procedures (Huynh & Meyer, 2010). The Robust z statistic was originally designed using the Rasch model (Huynh, 2000), but has been more recently applied to the 3PL and two-parameter partial credit models (Huynh & Meyer, 2010). The Robust z is essentially an outlier detection method and can be best conceptualized as a *z*-like statistic that is robust to outliers. As byproducts of the Robust z statistic, slope and intercept linking constants are also produced and have been found to be very similar to Stocking and Lord (1983) linking coefficients. Because the Robust z linking coefficients were not of primary interest in this study, they are provided in table format but are not discussed further.

The Robust z statistic is computed in two steps – the first focuses on the item discrimination estimate while the second step focuses on the item difficulty estimate. First, a univariate test of item *i*'s discrimination parameter estimate is computed as:

$$z_i = \frac{(D_{ai} - Md_{Da})}{0.74 * IQR_{Da}}. \quad (2)$$

Here, D_{ai} represents the difference between the natural logarithm of the old and new form discrimination (i.e., *a*) estimates or $D_{ai} = \ln(a_{iX}) - \ln(a_{iY})$, where X and Y represent the new and old form, respectively. It should be noted that new form item estimates have not yet been transformed to the old scale at this stage. Md_{Da} represents the median of the D_{ai} differences, taken over the common items and IQR_{Da} represents the interquartile range of the D_{ai} differences across all common items. The constant 0.74, when multiplied by the *IQR* emulates the standard deviation for a normal distribution. If the D_{ai} differences are normally distributed, the z_i statistic follows a normal (0,1) distribution and a critical value z^* may be specified. In this study, z^* was chosen to be 1.96 which represented a .05 significance level. Items with $| z_i | < 1.96$ were then used to compute the multiplicative scale transformation coefficient *A*, as:

$$A = \exp\left(\frac{\sum_{i=1}^{n_{ci,a}} D_{ai}}{n_{ci,a}}\right). \quad (3)$$

Here, $n_{ci,a}$ represents the number of common items with $| z_i | < 1.96$ in the previous equation. Next, differences are computed between the old and new form item difficulty estimates as:

$$D_{bi} = b_{yi} - A * b_{xi}. \quad (4)$$

Once, item difficulty differences are calculated for each of the n_{ci} items that remained after discarding items with unstable discrimination estimates, a univariate test to determine whether differences in item difficulty estimates were greater than expected due to random error, is computed as:

$$z_i = \frac{(D_{bi} - Md_{Db})}{0.74 * IQR_{Db}}. \quad (5)$$

Equation 5 is exactly the same as Equation 2, except that item discrimination is replaced with item difficulty. Again, the Robust z estimates obtained from Equation 5 were compared to the critical value z^* of ± 1.96 and if z_i exceeded z^*, the item was flagged for having unstable item difficulty estimates across forms, and was subsequently excluded from the computation of the additive scale transformation constant, *B*. The scaling coefficient *B*, is computed as:

$$B = \left[\frac{\sum_{i=1}^{n_{ci,ab}} D_{bi}}{n_{ci,ab}}\right], \quad (6)$$

where $n_{ci,ab}$ represents the number of common items that were found to have stable discrimination and difficulty estimates across forms. Several studies have found the Robust z linking coefficients to be very similar to the Stocking-Lord (1983) coefficients (Huynh & Meyer, 2010; Arce & Lau, 2011); however it was not our intent to compare scale linking methods in this study.

The two evaluation methods discussed up to this point have been model based and assume that the 3PL and GPC models fit the MC and FR item responses, reasonably well. As a result, ordinal logistic regression (OLR) was included as a third criterion since it makes no assumptions concerning model-data fit.

Ordinal logistic regression. OLR is an observed-score based procedure commonly used to test for the presence of uniform and non-uniform differential item functioning (DIF) in dichotomous and polytomous items (Elosua & Wells, 2013; Swaminathan & Rogers, 1990). Therefore, its inclusion enhances the repertoire of evaluative criteria by providing a different framework (i.e., DIF) within which to evaluate item performance.

In the context of IRT, DIF occurs when examinees of equal or similar abilities have different response probabilities for a particular item. In the current study, each FR item served as

the dependent variable, group membership was defined as either the old or new form (i.e., administration year), and ability was represented by common item total scores. Unlike the d^2 and Robust z procedures, OLR is model-free, meaning that its utility does not hinge on first properly fitting a measurement model.

In this study, OLR was used to test whether items from the same task model functioned differently across administrations. For each FR item, SAS software Version 9.4 was used to estimate two nested models: 1) a compact model where FR responses (scored 0-5) were predicted by total scores on the common items and 2) an augmented model which included the added predictors of group membership (i.e., form) and the interaction term between group membership and CI total scores. The difference between -2Log Likelihoods from both models was computed in order to carry out a likelihood ratio test, G_L^2. The G_L^2 test is distributed as a χ^2 with degrees of freedom equal to the difference in the number of parameters being estimated in both models (in our study, $df = 2$). The use of OLR as a DIF detection method has been shown to have inflated Type I error rates (Elosua & Wells, 2013; Swaminathan & Rogers, 1990) and as a result, a more stringent statistical significance level of $\alpha = 0.01$ was chosen.

The literature suggests that the most thorough way to test for DIF is to not only perform some type of statistical test, but to also include a measure of effect size to determine whether statistically significant results have any practical significance (Zumbo, 1999; Kim, Cohen, Alagoz, & Kim, 2007). The inclusion of an effect size measure is especially important with large sample sizes, such as those found in the current study. The effect size used in this study was demonstrated by Zumbo and Thomas (1997) and is essentially the change in R^2 values (i.e., R_Δ^2) associated with the family of nested models being compared. In order for an item to be flagged for DIF, Zumbo (1999) recommends that the G_L^2 test be statistically significant at the $\alpha = 0.01$ level and that R_Δ^2 be 0.13 or larger. The value of 0.13 corresponds to a medium effect under Cohen's (1988) effect size guidelines of 0.02, 0.13, and 0.26 corresponding to small, medium, and large effects. In our study, an effect size criterion of 0.02 was adopted for reasons discussed later.

Visual inspection. Scatter plots of the old versus new form item parameter estimates were created and compared against the results of the statistical flagging methods previously described in order to make sure flagging results were sensible. More specifically, inspection was limited to the discrimination estimate and global difficulty estimate associated with the GPC

model. Since the focus of this study was to determine whether FR items could be used interchangeably, the visual comparison of FR item estimates were made in the presence of the MC CI estimates. This was done so that the discrepancies between the MC CI estimates could be used as a rule of thumb for the degree of variation typically observed operationally. If task-model-derived FR items are to be treated as common items for equating purposes, it seems reasonable to expect that discrepancies between them should not be much larger than what is found between actual common items.

Results

Original and Pseudo-groups Data

Descriptive statistics for unweighted composite scores for both the original CINEG and pseudo-groups (i.e., RG) data, can be found in Tables 2 and 3, respectively. The sample sizes for the RG data were approximately half that of the original data; however this was necessary in order to create equivalent groups. Italian Language had the smallest sample sizes, yet they were considered large enough for the purpose of this study. The common item effect sizes for the original and RG data can be found in Table 4. The largest group differences associated with the original CINEG data were found between the 2012-2013 groups and 2012-2014 groups for both German and Italian. The ability differences between the RG samples were noticeably smaller; however in a few instances the goal of $ES \leq |0.01|$ could not be achieved by using the parental education matching variable (or any of the available matching variables for that matter). For these samples, effect sizes were still relatively small and always less than |0.05|.

Calibration and Linking

The calibrated item parameter estimates for the FR items are presented in Tables 5 and 6 for the original CINEG and RG data, respectively. The Stocking-Lord and Robust z scale transformation coefficients can be found in Table 7. It should be noted that no scale transformation was conducted with the RG datasets. Transformation coefficients are only provided because they were computed as a byproduct of the Robust z procedure. In general, the Stocking-Lord and Robust z procedures resulted in reasonably similar transformation coefficients; however the comparison of their similarities was not an objective of the current study.

Evaluation Criteria

In this section, evaluation results are organized by the AP subject tests to highlight similarities and differences between evaluation methods. Results for French Language are presented first and are followed by German Language and Italian Language. Because similar results were generally found with the CINEG and RG datasets, they are discussed together in order to avoid redundancy. Therefore, unless it is stated otherwise, it can be assumed that the following results apply to both the CINEG and RG datasets.

French Language.

Squared differences. The squared differences between the new and old form FR ICCs were summed over a standard normal theoretical posterior theta distribution and compared against the average differences observed in the operational CIs. Table 8 presents which of the French FR items were flagged due to having squared differences greater than two standard deviations above the mean. The third and fourth FR items, which each represent speaking tasks appeared to be consistently flagged.

Robust z. Table 9 presents flagging results using the Robust z method. Distinct flags are provided for distinguishing items with unstable *a*-parameter estimates (denoted as "A" in Table 9 and all subsequent Robust z tables) from those with unstable *b*-parameter estimates (denoted as "B" in Table 9 and all subsequent Robust z tables). However, it is possible for items to be flagged as having both unstable *a*- and *b*-parameter estimates. It should be noted that in this case, the *b*-parameter represents the global item difficulty parameter under the GPC model. For each FR item, the *a*- and *b*-parameter estimates are transformed to a "z-like" point estimate and were flagged as unstable if they exceed the critical *z* value of 1.96.

The item parameter estimates of the first FR item appeared to be most stable using the Robust z criteria and were never flagged. Whereas, the fourth FR item appeared to function most differently across forms and was flagged for having unstable *a*- and *b*-parameter estimates between the 2012-2013 and 2012-2014 links, in both datasets.

Ordinal logistic regression. Table 10 depicts the results of using OLR to detect differential item functioning (DIF) between forms. In order for an FR item to be flagged for DIF, it needed to have a statistically significant G_L^2 test statistic and an effect size ≥ 0.02. In both the CINEG and RG datasets, the fourth FR item was the only item flagged for DIF and was flagged in all instances.

Visual inspection. In order to check the reasonableness of the three previous flagging methods, scatter plots of the old versus the new form item estimates were created and can be found in Figures 1 and 2, for the CINEG and RG datasets, respectively. It should be noted that in the case of the CINEG data, the new form estimates have been transformed to the old form scale. Item parameter estimates of the FR items were plotted alongside those of the MC CIs in order to have a rough baseline comparison. More specifically, the differences observed between the MC CIs were used as reference points when judging whether differences between FR items were in line with what is typically seen between common items. If FR items built from the same task model showed similar discrepancies as the MC CIs, this was taken as evidence that they might be used interchangeably.

In both the CINEG and RG datasets, there was a higher agreement between the scatter plots and the Robust z method than between the scatter plots and the d^2 method. In general, similar trends were observed in the scatter plots, which provided some evidence that the flagging results appeared reasonable.

German Language.

Squared differences. Table 11 presents which of the German FR items were flagged according to the d^2 criterion. The fourth FR item (presentational speaking task) was flagged most frequently whereas the second (presentational writing) and third FR (interpersonal speaking) items were never flagged. The d^2 flagging results for German might be considered more stable than those found with French because there were no discrepancies found between the German CINEG and RG datasets.

Robust z. Table 12 presents flagging results for German Language using the Robust z method. For each FR item, the *a*- and *b*-parameter estimates are transformed to a "z-like" point estimate and are flagged as unstable if they exceed the critical *z* value of 1.96. The German flagging results using this method were similar to the d^2 flagging results in the sense that FR 4 was flagged in the 2012-2013 and 2012-2014 links in both datasets, which was more often than any other FR item. Based on the Robust z results, FR 2 appears to function most consistently across years, followed closely by FR 1.

Ordinal logistic regression. Table 13 depicts the results of using OLR to detect differential item functioning (DIF) between forms. In both the CINEG and RG datasets, the

fourth FR item was the only item flagged for DIF, but this difference was found in the 2012-2014 link. None of the FR items were flagged for DIF in the 2012-2013 or 2013-2015 links.

Visual inspection. Scatter plots of the old versus the new form item estimates for German Language can be found in Figures 3 and 4, for the CINEG and RG datasets, respectively. Again, it should be noted that for the CINEG data, the new form estimates have been transformed to the old form scale.

In both the CINEG and RG datasets, there was a higher agreement between the scatter plots and the Robust z method than between the scatter plots and the d^2 method. The only FR item parameter estimates that appeared to fall outside the cluster of the MC CI scatter points were the item discrimination estimates for FR 3 in the 2012-2013 link and FR2 in the 2012-2014 link.

Italian Language.

Squared differences. Table 14 depicts which of the Italian FR items were flagged according to the d^2 criterion that was previously described. In general, results were quite consistent across CINEG and RG datasets. Using the d^2 method, the second FR item (presentational writing task) exhibited the greatest differences across forms and was flagged 100% of the time. The first (interpersonal writing) and fourth FR items (presentational speaking) were also flagged very frequently; however the third FR item (interpersonal speaking) showed the greatest stability and was never flagged.

Robust z. The Robust z flagging results for Italian Language are presented in Table 15. The flagging results differed somewhat between the CINEG or RG datasets; however there was a general tendency for FR 3 to be flagged least often. This finding agreed with the flagging results found using the d^2 method.

Ordinal logistic regression. There were noticeably more flags found for Italian Language using OLR than for the other two subjects (see Table 16). However, the flagging results using OLR agreed with the d^2 and Robust z results such that FR 3 appeared most stable across forms (i.e., was never flagged). There was also a high degree of consistency in the flags found between the CINEG and RG datasets, with only one flagging discrepancy found in the 2013-2015 link.

Visual inspection. The scatter plots of the new versus old form *a*- and *b*-parameter estimates for the MC CIs and four FR items can be found in Figures 5 and 6 for the CINEG and RG datasets, respectively. Before considering the similarities and differences between the FR

items it should be noted that there were several instances where discrepancies between the MC CIs were larger than expected. More specifically, there were several MC CIs in the 2012-2014 link that had noticeably large differences in their *a*- and *b*-parameter estimates across forms. Similarly, several of the 2013-2015 MC CIs had *a*-parameters that appeared to shift quite considerably between the two years. These noticeable differences in the CI estimates not only could have distorted the visual inspection of the FR item estimates but could have also influenced the flagging results presented earlier. Limitations related to this issue will be revisited in the discussion section.

The visual inspection of the *a*- and *b*-item parameter estimates showed that there tended to be fewer FR items with noticeable discrepancies between the 2013 and 2015 forms. The biggest differences were found with the FR 1 and FR 4 *b*-parameter estimates between the 2012 and 2013 forms.

Discussion

The purposes of this study were to a) demonstrate how to evaluate task-model-derived FR items based on methods originally designed to evaluate MC items and b) examine whether task-model-derived FR items behaved similarly enough across administrations in order to be treated interchangeably for equating purposes in the future. In this study, FR items generated from a particular task model were regarded as isomorphs, meaning they were written with the goal of being mutually interchangeable with respect to content and psychometric properties (Bejar, 2002). Therefore, methods traditionally used to detect unstable common items and differential item functioning were used to evaluate whether FR items built from the same task model could actually be treated as isomorphs. In order to do this, items belonging to the same task model were treated as though they were the same exact item for statistical purposes and were evaluated using the same set of criteria that are traditionally applied to detect instability in MC common items.

Even though common-item sets should theoretically resemble a mini version of the full length test, they are routinely composed of only MC items – even when the test is mixed-format. It naturally follows then, that the majority (if not all) of the methods identified in the literature for detecting unstable common items were developed with only MC items in mind. Therefore, in order to use the methods that resembled some form of outlier detection (i.e., the Robust z and d^2), first the score scale of the FR items had to be transformed to the MC score scale (0/1). In

order to do this, FR scores were divided by the maximum number of score points, which resulted in a new score scale that ranged from 0 to 1. While this transformation may have resulted in the loss of information, without it, the majority of detection methods could not have been used. Scatter plots of the old form versus new form FR item parameter estimates were visually inspected in order to help safeguard against interpreting unreasonable results that may have stemmed from this transformation. In general, the flagging results agreed with the discrepancies observed in the scatter plots, which implies the transformation likely did not introduce enough systematic error to render results untrustworthy.

For each of the three subject areas, the flagging methods generally appeared to agree with one another, which provided some evidence of their utility. For example, for Italian Language, the third FR item was routinely flagged least frequently by the Robust z, d^2, and *OLR* methods and this was found across all three links (i.e., 2012-2013, 2012-2014, and 2013-2015). Similarly, for French Language, FR 1 was routinely flagged less than any of the other FR items and FR 4 was flagged noticeably most often. For German Language, FR 2 appeared to behave most consistently across forms and was flagged the least number of times by each method. Therefore, at the very least, there were consistent flagging patterns seen across the three chosen detection methods. These patterns were also generally supported by the visual inspection of the old form versus new form item parameter estimates. Therefore, if one were to use the results of this study to help inform the choice of which FR items might perform best as common items for equating purposes, one would be inclined to select the following: FR 1 for French Language tests, FR 2 for German Language, and FR 3 for Italian Language. It was interesting to find that the particular type of task (i.e., interpersonal writing, presentational writing, etc.) did not appear to consistently outperform others. For example, FR 1 always corresponded to an interpersonal writing task, FR 2 always corresponded to a presentational writing task, FR 3 always corresponded to a culmination of five short interpersonal speaking tasks, and FR 4 always corresponded to a presentational speaking task. Yet, when you compare results across tests, either FR 1, FR 2, or FR 3 appeared to function most similarly across years, depending on the subject. This finding could reflect the case where certain task models are more robust depending on the subject; however additional research would be needed to draw this conclusion. Perhaps the strongest conclusion that can be drawn from this study is that FR 4 is likely the worst item to consider for use as a common item. This finding could be related to its content (i.e., presentational speaking)

and the variability in scoring or it could be reflective of its position as the last item on the test. It is very reasonable to expect that fatigue may set it by this point and that examinee ability and performance may not be accurately captured by this item.

The flagging results were generally very consistent between the original CINEG and pseudo-group RG datasets, which implies that the employed methods might be quite stable and may not be very affected by group differences. The largest group difference (i.e., $ES = 0.236$) among the original CINEG datasets was found between the 2012 and 2014 examinees for German Language. According to Kolen and Brennan (2014) group differences of 0.30 standard deviation units are typically needed in order to see differences in *equating* methods [emphasis added]. It is unknown whether this rule of thumb can be applied to the detection methods used in this study; however one might conclude that the original data did not have large enough group differences to necessitate the creation of the RG datasets.

Before closing, it is worth discussing several caveats regarding the alternate ways in which the flagging methods could have been used in this study. First, there was no iterative purification of the operational MC common-item set, except for the Robust z procedure. However, for the Robust z procedure the purification only affects the computation of the scaling constants. In each subject area, there were a few instances where some of the operational MC common items appeared to function quite differently across forms. However, because the focus of this study was on detection in the FR item set, flagging results for the MC items were not presented. If the poor performing common items would have been removed from the group of items used to create the criterion for FR flagging purposes, the results might have been different. For the d^2 statistic, this change could have resulted in the FR items being flagged more often because the comparative criteria would be more stringent (i.e., based on common items that behaved more similarly across forms). Purification of the common-item set would probably have less of an impact on the Robust z method since it was designed to be robust to outliers.

A second caveat is that the use of different critical values for the Robust z and OLR methods could have led to different results. For the Robust z method, a critical value of $z^* = \pm 1.96$ was employed, which corresponds to nominal type one error rate of 5%. However, Agresti and Finlay (2009) suggest that an observation is an outlier if it falls more than 1.5(IQR) above the upper quartile or 1.5(IQR) below the lower quartile, which corresponds to a critical value of $z^* = \pm 2.7$. Therefore, if this more relaxed criteria of $z^* = 2.7$ were used, it is very likely that

fewer FR items would have been flagged. Similarly, with the OLR method, a nominal type one error rate of 1% was used due to the large sample sizes and the critical effect size was set at 0.02. These critical values also could have been set to more rigorous or more lenient values. For example, Zumbo (1999) recommends that the G_L^2 test be statistically significant at the α = 0.01 level and that R_Δ^2 be 0.13 or larger in order for an item to be flagged for DIF. Even though we adopted his recommendation for the α level, the effect size guidelines of 0.13 seemed too lenient to produce useful results. Actually, if the effect size guideline of requiring R_Δ^2 to be at least 0.13 were adopted, only FR 1 from the Italian 2012-2013 link would have been flagged. This did not appear reasonable given the results of the other flagging methods as well as the visual inspection of item parameter estimates. Kim et al. (2007) reported similar findings when investigating DIF in polytomous items, such that none of their studied items approached an R_Δ^2 of 0.13. Flagging methods for DIF or otherwise, are usually employed as a first step in a longer post-administration item analysis process, therefore, a method has very little utility if it essentially flags no items.

In this study, methods that were originally designed to detect DIF and unstable MC common items were adapted to evaluate the robustness of four task models and the FR items developed from each. These methods were applied with the intent that FR items with favorable results might be treated as common items for equating purposes in the future. In general, results were quite similar across methods and were useful in identifying at least one FR item in each subject area that if included in the common-item set, would improve the content representativeness of the CI set and thereby hopefully improve equating accuracy. It naturally follows that the next step is to actually treat these identified FR items as common and evaluate the extent to which they impact equating results.

References

Agresti, A. & Finlay, B. (2009). Statistical Methods for the Social Sciences, 4th edition. Upper Saddle River, NJ: Prentice Hall.

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.

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.

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

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

Elosua, P. & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT, and ordinal logistic regression. *Psicologica, 34*, 327-342.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*(4), 358-377.

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology, 14*, 1-27.

Huynh, H. (2000, June). *Guidelines for Rasch Linking for PACT*. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H. & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15(2). Available online: http://pareonline.net/getvn.asp?v=15&n=2.

Kim, S.-H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116.

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

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*. Retrieved June 8th, 2016 from www.psych.umn.edu/psylabs/CATCentral/

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.

Pearson, Inc. (2012). *2011-2012 Maryland School Assessment (MSA) Science Annual Technical Report: Grades 5 and 8.* Retrieved from http://www.marylandpublicschools.org/msde/divisions/planningresultstest/2012+MSA+Science+Technical+Report.htm

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

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R.L. Brennan (Ed.) *Educational Measurement*, 4th Edition (307-353). Washington, DC: American Council on Education.

Stocking, M., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207-210.

Swaminathan, H. & Rogers J.H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Zumbo, B. D., & Thomas, D. R. (1997) *A measure of effect size for a model-based approach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Table 1

Description of Matching Variable

Code	Highest Level of Education Achieved by Either Parent
0	No Response
1	Grade School
2	Some High School
3	High School Diploma
4	Business School
5	Some College
6	Associates Degree
7	Bachelors Degree
8	Some Graduate
9	Graduate Degree

Table 2

Descriptive Statistics for the Original CINEG Data

Subject	*Year*	*N*	*Mean*	*SD*	*Skew.*	*Kurt.*	*Min.*	*Max.*
French								
	2012	17,006	55.651	14.755	-0.376	-0.535	1	85
	2013	20,000	55.995	14.733	-0.436	-0.522	0	85
	2014	20,000	54.985	14.300	-0.287	-0.447	5	85
	2015	20,000	51.342	14.692	-0.075	-0.707	3	85
German								
	2012	2,157	57.935	16.696	-0.349	-0.756	8	85
	2013	4,803	56.049	15.983	-0.116	-0.743	9	85
	2014	4,867	58.008	16.448	-0.286	-0.775	5	85
	2015	4,885	54.357	17.064	-0.024	-0.894	9	85
Italian								
	2012	1,760	58.991	16.889	-0.200	-0.664	13	90
	2013	1,905	56.912	18.115	-0.043	-0.941	13	90
	2014	2,292	52.435	17.241	0.106	-0.798	8	88
	2015	2,520	52.278	18.500	0.151	-0.951	5	90

Note. Statistics represent unweighted composite scores.

Table 3

Descriptive Statistics for the Random Groups Data

Subject	*Form Y*	*Form X*	*N*	*Mean*	*SD*	*Skew.*	*Kurt.*	*Min.*	*Max.*
French									
	2012	2013	9,689	55.714	14.604	-0.387	-0.513	5	85
		2014	9,683	55.492	14.733	-0.377	-0.536	5	85
	2013	2012	9,689	55.783	14.763	-0.430	-0.517	0	85
		2015	4,878	57.196	14.142	-0.485	-0.415	11	84
	2014	2012	9,683	54.948	14.261	-0.287	-0.433	5	85
	2015	2013	4,878	53.141	14.051	-0.144	-0.623	10	84
German									
	2012	2013	2,157	57.935	16.696	-0.349	-0.756	8	85
		2014	2,061	57.746	16.609	-0.327	-0.760	8	85
	2013	2012	2,157	58.093	16.014	-0.209	-0.803	14	85
		2015	3,000	57.611	15.235	-0.164	-0.707	15	85
	2014	2012	2,061	61.153	16.089	-0.496	-0.600	12	85
	2015	2013	3,000	56.260	16.347	-0.086	-0.876	9	85
Italian									
	2012	2013	1,150	58.618	16.957	-0.153	-0.678	13	90
		2014	1,400	59.106	16.808	-0.195	-0.660	13	90
	2013	2012	1,150	58.143	18.232	-0.098	-0.914	13	90
		2015	1,450	56.463	17.937	-0.022	-0.919	13	90
	2014	2012	1,400	54.455	17.002	0.032	-0.834	8	88
	2015	2013	1,450	52.125	18.378	0.190	-0.943	5	90

Note. Statistics represent unweighted composite scores.

Table 4

Effect Size of Group Differences Using Common Item Scores

Data Source	**Old - New Group**	**German**	**French**	**Italian**
CINEG				
	2012-2013	0.148	-0.017	0.121
	2012-2014	0.236	0.019	0.136
	2013-2015	0.021	0.084	0.033
Random Groups				
	2012-2013	0.032	0.001	0.019
	2012-2014	0.011	0.001	0.009
	2013-2015	0.015	0.045	0.000

Note. Positive values indicate old group is higher achieving.

Table 5

Item Parameter Estimates for the CINEG Data

	Item	*a*	*b*	d_1	d_2	d_3	d_4	d_5	d_6
German									
2012	**FR1**	1.137	-0.981	0	1.856	0.632	-0.030	-0.826	-1.632
	FR2	1.284	-0.889	0	0.964	1.138	0.296	-0.917	-1.480
	FR3	0.791	-1.064	0	1.046	0.727	0.065	-0.730	-1.108
	FR4	0.359	-0.385	0	0.367	0.645	0.432	-1.009	-0.436
2013									
	FR1	1.261	-1.366	0	1.949	1.008	-0.101	-1.172	-1.684
	FR2	1.121	-0.792	0	1.145	1.030	0.201	-0.811	-1.566
	FR3	1.290	-0.948	0	1.425	0.741	0.077	-0.884	-1.358
	FR4	0.554	-0.541	0	0.172	0.521	0.860	-0.581	-0.973
2014									
	FR1	1.145	-1.250	0	1.844	1.099	-0.333	-0.892	-1.719
	FR2	0.946	-1.216	0	1.138	1.150	0.380	-0.968	-1.701
	FR3	0.853	-1.377	0	1.444	0.796	0.011	-0.788	-1.463
	FR4	0.524	-0.853	0	0.280	0.953	0.426	-0.561	-1.099
2015									
	FR1	1.367	-1.262	0	1.708	0.951	-0.021	-1.035	-1.603
	FR2	1.087	-0.882	0	1.325	0.931	0.201	-1.010	-1.447
	FR3	1.205	-0.957	0	1.368	0.641	0.086	-0.816	-1.279
	FR4	0.581	-0.717	0	0.186	0.824	0.621	-0.583	-1.048
French									
2012									
	FR1	0.768	-0.972	0	2.945	1.484	-0.122	-1.721	-2.586
	FR2	0.735	-0.573	0	1.745	1.225	0.130	-1.166	-1.934
	FR3	0.612	-0.531	0	1.629	0.803	0.111	-1.031	-1.512
	FR4	0.305	-0.122	0	1.458	0.879	0.224	-1.107	-1.454
2013									
	FR1	0.705	-0.945	0	2.668	1.834	-0.365	-1.590	-2.548
	FR2	0.777	-0.759	0	1.883	1.732	0.098	-1.460	-2.254
	FR3	0.578	-0.781	0	1.347	0.859	0.136	-0.973	-1.369
	FR4	0.384	-0.642	0	1.525	1.235	0.318	-1.317	-1.761
2014									
	FR1	0.848	-0.837	0	2.685	0.953	-0.245	-1.326	-2.067
	FR2	0.940	-0.749	0	1.923	1.286	0.030	-1.232	-2.007
	FR3	0.773	-0.770	0	2.036	0.705	-0.070	-1.142	-1.530
	FR4	0.443	-0.727	0	2.199	0.265	0.517	-1.312	-1.669
2015									
	FR1	0.838	-0.996	0	2.410	1.260	-0.071	-1.452	-2.146
	FR2	0.847	-0.824	0	1.638	1.506	0.182	-1.254	-2.072
	FR3	0.706	-0.648	0	2.151	0.644	-0.388	-1.112	-1.295
	FR4	0.384	-0.417	0	2.586	0.947	0.273	-1.627	-2.178

Italian									
2012									
	FR1	0.949	-0.341	0	1.963	0.780	0.050	-1.251	-1.542
	FR2	0.689	-0.280	0	1.870	0.736	-0.092	-1.329	-1.184
	FR3	0.902	-0.669	0	1.324	0.777	-0.044	-0.781	-1.276
	FR4	0.476	-0.278	0	1.835	0.489	-0.030	-0.845	-1.450
2013									
	FR1	0.758	-1.330	0	1.339	1.277	-0.048	-1.187	-1.381
	FR2	0.998	-0.703	0	1.282	0.958	-0.107	-0.931	-1.202
	FR3	1.037	-0.710	0	1.393	0.566	0.073	-0.721	-1.310
	FR4	0.551	0.368	0	-0.655	0.956	0.838	-0.378	-0.762
2014									
	FR1	0.640	-0.408	0	1.280	0.518	-0.033	-0.814	-0.951
	FR2	0.949	-0.774	0	1.117	0.938	0.052	-0.926	-1.180
	FR3	1.189	-0.856	0	1.583	0.516	-0.141	-0.754	-1.203
	FR4	0.573	-0.482	0	0.905	0.659	0.607	-0.973	-1.199
2015									
	FR1	0.936	-0.883	0	2.060	0.855	-0.156	-1.242	-1.518
	FR2	0.905	-0.411	0	1.329	0.486	0.044	-0.774	-1.085
	FR3	1.015	-0.840	0	1.307	0.578	0.086	-0.760	-1.211
	FR4	0.520	0.078	0	1.076	0.658	-0.337	-1.060	-0.337

Table 6

Item Parameter Estimates for the Random Groups Data

Year	**Use**	**Item**	***a***	***b***	d_1	d_2	d_3	d_4	d_5	d_6
German										
2012	Link 13	FR1	1.137	-0.981	0	1.856	0.632	-0.030	-0.826	-1.632
	Link 13	FR2	1.284	-0.889	0	0.964	1.138	0.296	-0.917	-1.480
	Link 13	FR3	0.791	-1.064	0	1.046	0.727	0.065	-0.730	-1.108
	Link 13	FR4	0.359	-0.385	0	0.367	0.645	0.432	-1.009	-0.436
	Link 14	FR1	1.136	-0.963	0	1.845	0.630	-0.022	-0.828	-1.625
	Link 14	FR2	1.285	-0.887	0	0.993	1.130	0.289	-0.927	-1.485
	Link 14	FR3	0.783	-1.046	0	1.017	0.747	0.072	-0.723	-1.113
	Link 14	FR4	0.368	-0.386	0	0.364	0.684	0.396	-0.994	-0.451
2013										
	New	FR1	1.173	-1.405	0	1.980	1.099	-0.053	-1.185	-1.841
	New	FR2	0.999	-0.817	0	1.288	1.147	0.194	-0.958	-1.670
	New	FR3	1.143	-1.081	0	1.730	0.754	0.029	-0.994	-1.518
	New	FR4	0.505	-0.508	0	0.119	0.479	0.956	-0.521	-1.032
	Link 15	FR1	1.262	-1.337	0	1.841	1.022	-0.049	-1.102	-1.712
	Link 15	FR2	1.075	-0.791	0	1.198	1.066	0.181	-0.891	-1.553
	Link 15	FR3	1.229	-1.036	0	1.609	0.701	0.027	-0.924	-1.412
	Link 15	FR4	0.543	-0.503	0	0.110	0.446	0.889	-0.484	-0.960

2014										
	New	FR1	1.227	-1.083	0	1.617	1.014	-0.255	-0.786	-1.591
	New	FR2	0.984	-1.129	0	0.912	1.229	0.387	-0.885	-1.643
	New	FR3	0.893	-1.254	0	1.173	0.792	0.108	-0.793	-1.280
	New	FR4	0.530	-0.783	0	0.415	0.681	0.426	-0.461	-1.060
2015										
	New	FR1	1.215	-1.193	0	1.903	1.056	-0.013	-1.159	-1.788
	New	FR2	0.977	-0.770	0	1.466	1.045	0.221	-1.119	-1.612
	New	FR3	1.082	-0.851	0	1.521	0.723	0.104	-0.906	-1.443
	New	FR4	0.532	-0.586	0	0.221	0.917	0.696	-0.641	-1.193
French										
2012	Link 13	FR1	0.752	-0.951	0	2.849	1.551	-0.091	-1.701	-2.608
	Link 13	FR2	0.730	-0.577	0	1.717	1.258	0.130	-1.142	-1.963
	Link 13	FR3	0.606	-0.530	0	1.690	0.827	0.101	-1.052	-1.566
	Link 13	FR4	0.299	-0.120	0	1.495	0.964	0.216	-1.082	-1.594
	Link 14	FR1	0.764	-0.974	0	2.924	1.524	-0.129	-1.713	-2.606
	Link 14	FR2	0.724	-0.570	0	1.767	1.240	0.144	-1.178	-1.973
	Link 14	FR3	0.613	-0.504	0	1.644	0.819	0.129	-1.041	-1.550
	Link 14	FR4	0.300	-0.118	0	1.508	0.960	0.137	-1.058	-1.547
2013										
	New	FR1	0.737	-0.928	0	2.661	1.752	-0.386	-1.551	-2.476
	New	FR2	0.797	-0.736	0	1.824	1.691	0.097	-1.408	-2.204
	New	FR3	0.594	-0.748	0	1.271	0.875	0.123	-0.936	-1.332
	New	FR4	0.410	-0.616	0	1.460	1.231	0.315	-1.289	-1.718
	Link 15	FR1	0.748	-0.905	0	2.620	1.725	-0.380	-1.527	-2.438
	Link 15	FR2	0.809	-0.716	0	1.796	1.665	0.096	-1.386	-2.170
	Link 15	FR3	0.603	-0.728	0	1.251	0.861	0.121	-0.922	-1.311
	Link 15	FR4	0.417	-0.598	0	1.438	1.212	0.310	-1.269	-1.692
2014										
	New	FR1	0.808	-0.862	0	2.905	0.963	-0.266	-1.408	-2.194
	New	FR2	0.895	-0.774	0	2.067	1.345	0.039	-1.325	-2.126
	New	FR3	0.744	-0.756	0	2.117	0.744	-0.062	-1.183	-1.617
	New	FR4	0.425	-0.725	0	2.332	0.326	0.520	-1.361	-1.817
2015										
	New	FR1	0.752	-1.083	0	2.313	1.587	-0.045	-1.521	-2.334
	New	FR2	0.757	-0.949	0	1.652	1.714	0.301	-1.415	-2.253
	New	FR3	0.683	-0.685	0	2.203	0.724	-0.420	-1.159	-1.348
	New	FR4	0.345	-0.547	0	2.745	1.159	0.256	-1.773	-2.387

Italian										
2012	Link 13	FR1	0.959	-0.329	0	2.000	0.719	0.082	-1.229	-1.572
	Link 13	FR2	0.747	-0.268	0	1.794	0.732	-0.016	-1.369	-1.141
	Link 13	FR3	0.932	-0.623	0	1.325	0.726	-0.011	-0.781	-1.260
	Link 13	FR4	0.511	-0.324	0	1.940	0.423	0.021	-0.910	-1.473
	Link 14	FR1	0.930	-0.356	0	1.982	0.781	0.048	-1.273	-1.538
	Link 14	FR2	0.684	-0.271	0	1.858	0.776	-0.143	-1.284	-1.207
	Link 14	FR3	0.908	-0.667	0	1.365	0.763	-0.019	-0.802	-1.307
	Link 14	FR4	0.467	-0.299	0	1.841	0.494	-0.060	-0.879	-1.395
2013										
	New	FR1	0.702	-1.321	0	1.224	1.358	0.027	-1.280	-1.330
	New	FR2	1.011	-0.682	0	1.229	1.009	-0.107	-0.955	-1.176
	New	FR3	1.068	-0.645	0	1.474	0.549	0.021	-0.766	-1.278
	New	FR4	0.513	0.428	0	0.877	1.161	0.668	-0.179	-0.774
	Link 15	FR1	0.703	-1.326	0	1.222	1.356	0.027	-1.278	-1.328
	Link 15	FR2	1.013	-0.688	0	1.228	1.007	-0.107	-0.954	-1.174
	Link 15	FR3	1.069	-0.651	0	1.472	0.548	0.021	-0.765	-1.276
	Link 15	FR4	0.514	0.420	0	0.875	1.159	0.667	-0.178	-0.773
2014										
	New	FR1	0.641	-0.420	0	1.175	0.561	-0.017	-0.786	-0.933
	New	FR2	0.955	-0.817	0	1.090	0.887	0.081	-0.946	-1.112
	New	FR3	1.303	-0.851	0	1.526	0.457	-0.093	-0.729	-1.162
	New	FR4	0.565	-0.543	0	0.923	0.653	0.549	-1.000	-1.125
2015										
	New	FR1	0.933	-0.790	0	2.246	0.823	-0.173	-1.351	-1.546
	New	FR2	0.792	-0.220	0	1.324	0.559	0.033	-0.798	-1.117
	New	FR3	1.003	-0.700	0	1.315	0.572	0.097	-0.752	-1.232
	New	FR4	0.518	0.233	0	1.106	0.709	-0.450	-1.040	-0.325

Table 7
Scale Transformation Coefficients

	CINEG Data				**RG Data**			
	Stocking Lord		**Robust z**		**Stocking Lord**		**Robust z**	
	A	***B***	***A***	***B***	***A***	***B***	***A***	***B***
German								
2012-2013	0.932	-0.087	1.016	-0.089	0.930	-0.031	1.081	-0.070
2012-2014	0.991	-0.292	1.110	-0.281	1.012	0.021	1.030	-0.029
2013-2015	0.901	-0.175	1.053	-0.242	0.894	-0.189	1.116	-0.212
French								
2012-2013	1.022	0.020	0.988	0.004	0.985	0.009	0.993	-0.009
2012-2014	0.953	-0.024	0.998	-0.045	0.967	-0.003	0.989	-0.002
2013-2015	0.988	-0.076	1.014	-0.064	0.892	0.046	1.087	0.116
Italian								
2012-2013	0.974	-0.119	1.016	-0.091	0.998	-0.007	1.007	-0.021
2012-2014	1.056	-0.053	0.938	-0.138	0.996	0.074	0.943	-0.004
2013-2015	0.988	-0.150	0.989	-0.191	0.960	-0.097	0.966	-0.227

Table 8

Flagging Results Using the d^2 Statistic for French Language

		d^2 Flags		
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013		Flagged	Flagged	Flagged
2012-2014	Flagged		Flagged	Flagged
2013-2015	Flagged		Flagged	Flagged
RG Design				
2012-2013				
2012-2014			Flagged	Flagged
2013-2015			Flagged	Flagged

Note. d^2 criterion is defined as a d^2 value two standard deviations above the mean d^2 for the operational MC common items. One common item in the 2012-2013 RG data had such a large d^2 value that if removed, FR 3 and FR 4 would have been flagged.

Table 9

French Language Flagging Results Using the Robust z Procedure

		Robust z Flags		
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013		B	B	A,B
2012-2014		A	A	A,B
2013-2015			A	
RG Design				
2012-2013			B	A,B
2012-2014		A	B	A,B
2013-2015				

Note. "A" and "B" indicate item was flagged as having unstable discrimination and/or global location estimates, respectively.

Table 10

French Language Flagging Results Using Ordinal Logistic Regression

		OLR Flags		
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013				Flagged
2012-2014				Flagged
2013-2015				Flagged
RG Design				
2012-2013				Flagged
2012-2014				Flagged
2013-2015				Flagged

Note. Flagged items had a significant G_L^2 test statistic ($p < .01$) and an effect size ≥ 0.02.

Table 11

Flagging Results Using the d^2 Statistic for German Language

	d^2 Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Data				
2012-2013	Flagged			Flagged
2012-2014				Flagged
2013-2015				
RG Data				
2012-2013	Flagged			Flagged
2012-2014				Flagged
2013-2015				

Note. d^2 criterion is defined as a d^2 value two standard deviations above the mean d^2 for the operational MC common items.

Table 12

German Language Flagging Results Using the Robust z Procedure

	Robust z Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013	B		A	A
2012-2014				B
2013-2015				
RG Design				
2012-2013	B		A	A
2012-2014		B	B	B
2013-2015				

Note. "A" and "B" indicate item was flagged as having unstable discrimination and/or global location estimates, respectively.

Table 13

German Language Flagging Results Using Ordinal Logistic Regression

	OLR Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013				
2012-2014				Flagged
2013-2015				
RG Design				
2012-2013				
2012-2014				Flagged
2013-2015				

Note. Flagged items had a significant G_L^2 test statistic ($p < .01$) and an effect size ≥ 0.02.

Table 14

Flagging Results Using the d^2 Statistic for Italian Language

	d^2 Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013	Flagged	Flagged		Flagged
2012-2014	Flagged	Flagged		Flagged
2013-2015	Flagged	Flagged		Flagged
RG Design				
2012-2013	Flagged	Flagged		
2012-2014		Flagged		Flagged
2013-2015	Flagged	Flagged		Flagged

Note. d^2 criterion is defined as a d^2 value two standard deviations above the mean d^2 for the operational MC common items.

Table 15

Italian Language Flagging Results Using the Robust z Procedure

	Robust z Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013	B	A,B		B
2012-2014	A	A		
2013-2015	B			
RG Design				
2012-2013	B			B
2012-2014	A	A,B	A	
2013-2015				B

Note. "A" and "B" indicate item was flagged as having unstable discrimination and/or global location estimates, respectively.

Table 16

Italian Language Flagging Results Using Ordinal Logistic Regression

	OLR Flags			
Source	**FR 1**	**FR 2**	**FR 3**	**FR 4**
CINEG Design				
2012-2013	Flagged	Flagged		Flagged
2012-2014		Flagged		
2013-2015	Flagged			Flagged
RG Design				
2012-2013	Flagged	Flagged		Flagged
2012-2014		Flagged		
2013-2015	Flagged	Flagged		Flagged

Note. Flagged items had a significant G_L^2 test statistic ($p < .01$) and an effect size ≥ 0.02.

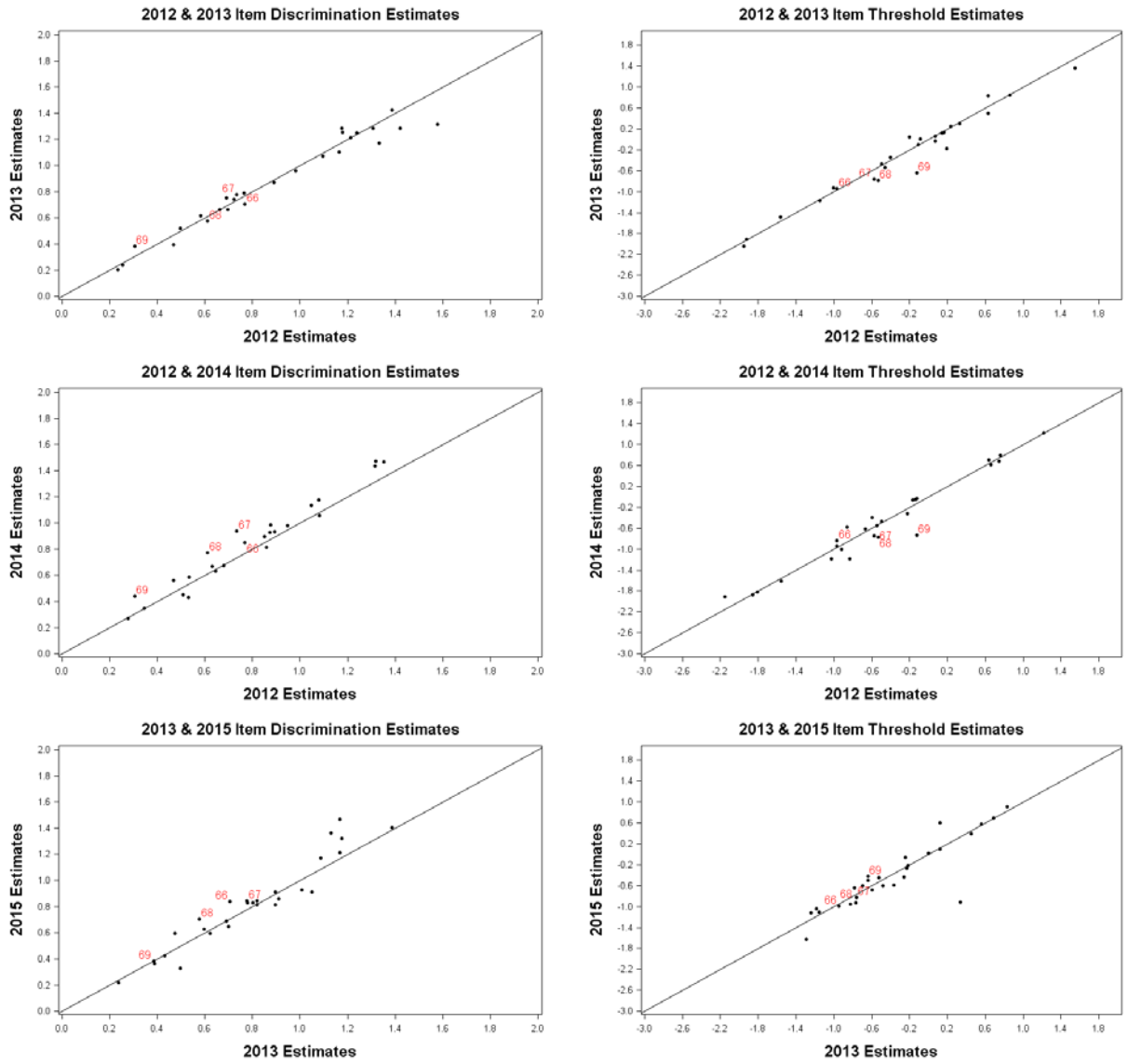

Figure 1. Visual inspection of old versus new form item parameter estimates for French Language using the original CINEG data.

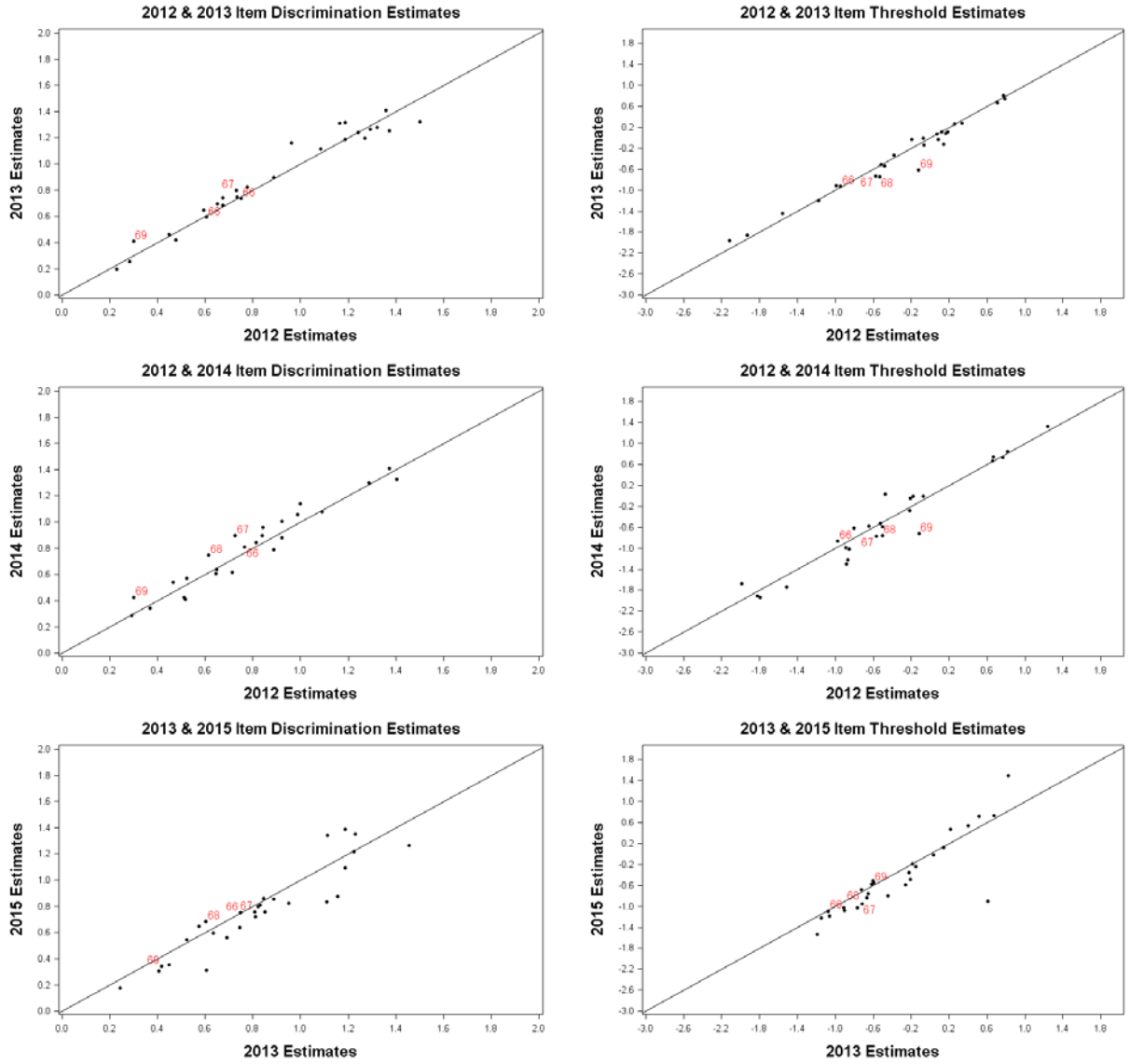

Figure 2. Visual inspection of old versus new form item parameter estimates for French Language using the RG datasets.

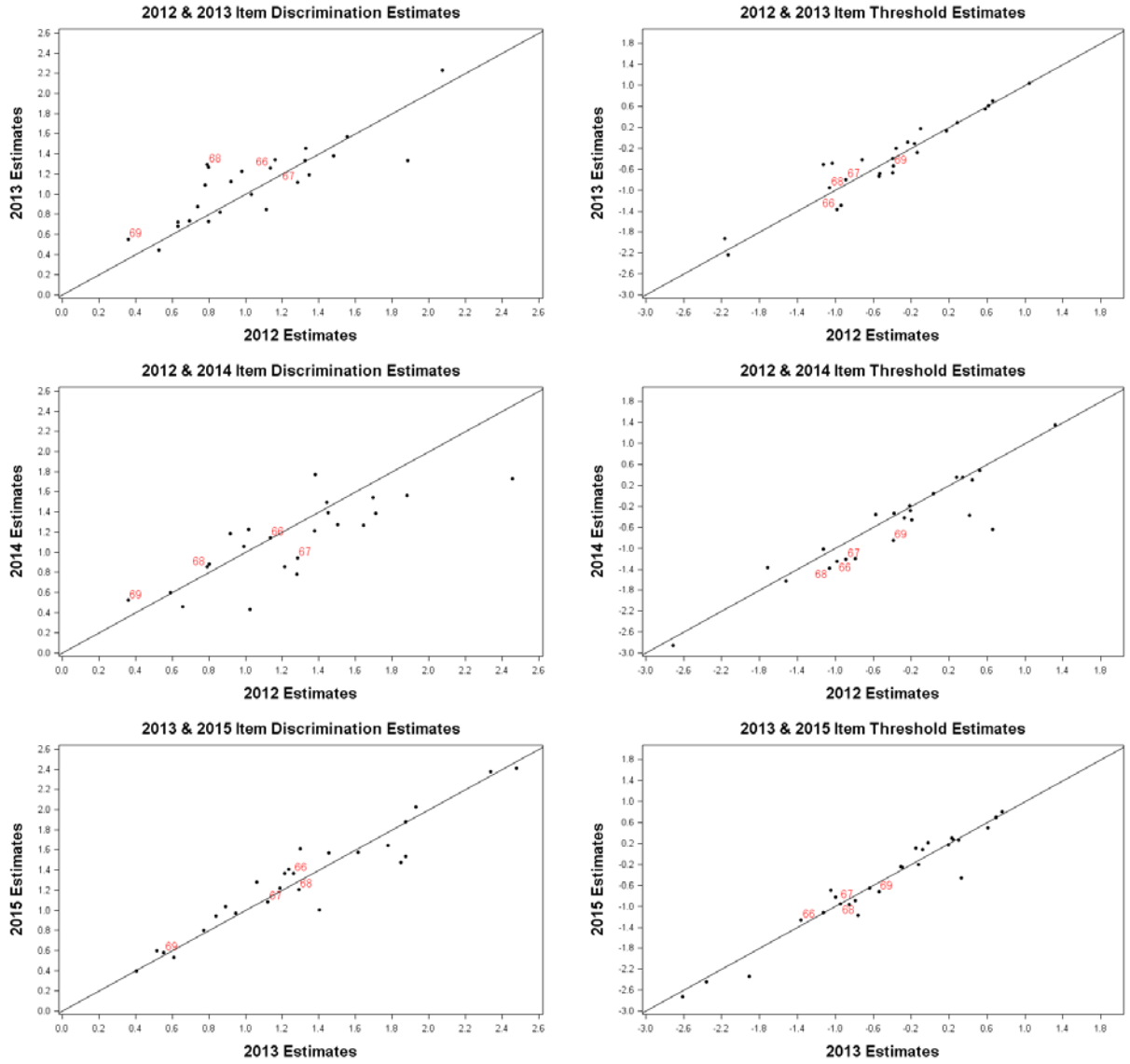

Figure 3. Visual inspection of old versus new form item parameter estimates for German Language using the original CINEG data.

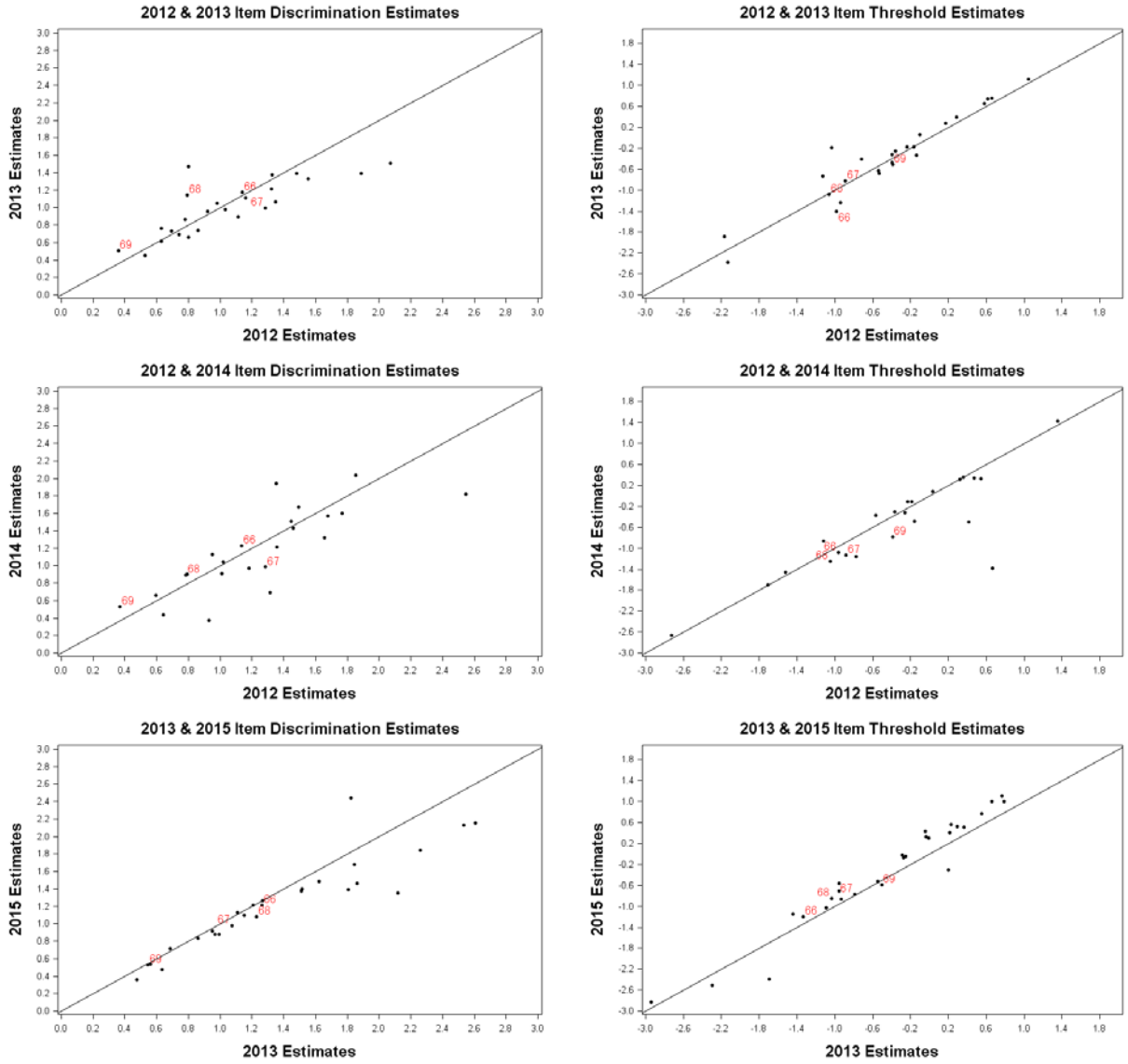

Figure 4. Visual inspection of old versus new form item parameter estimates for German Language using the RG datasets.

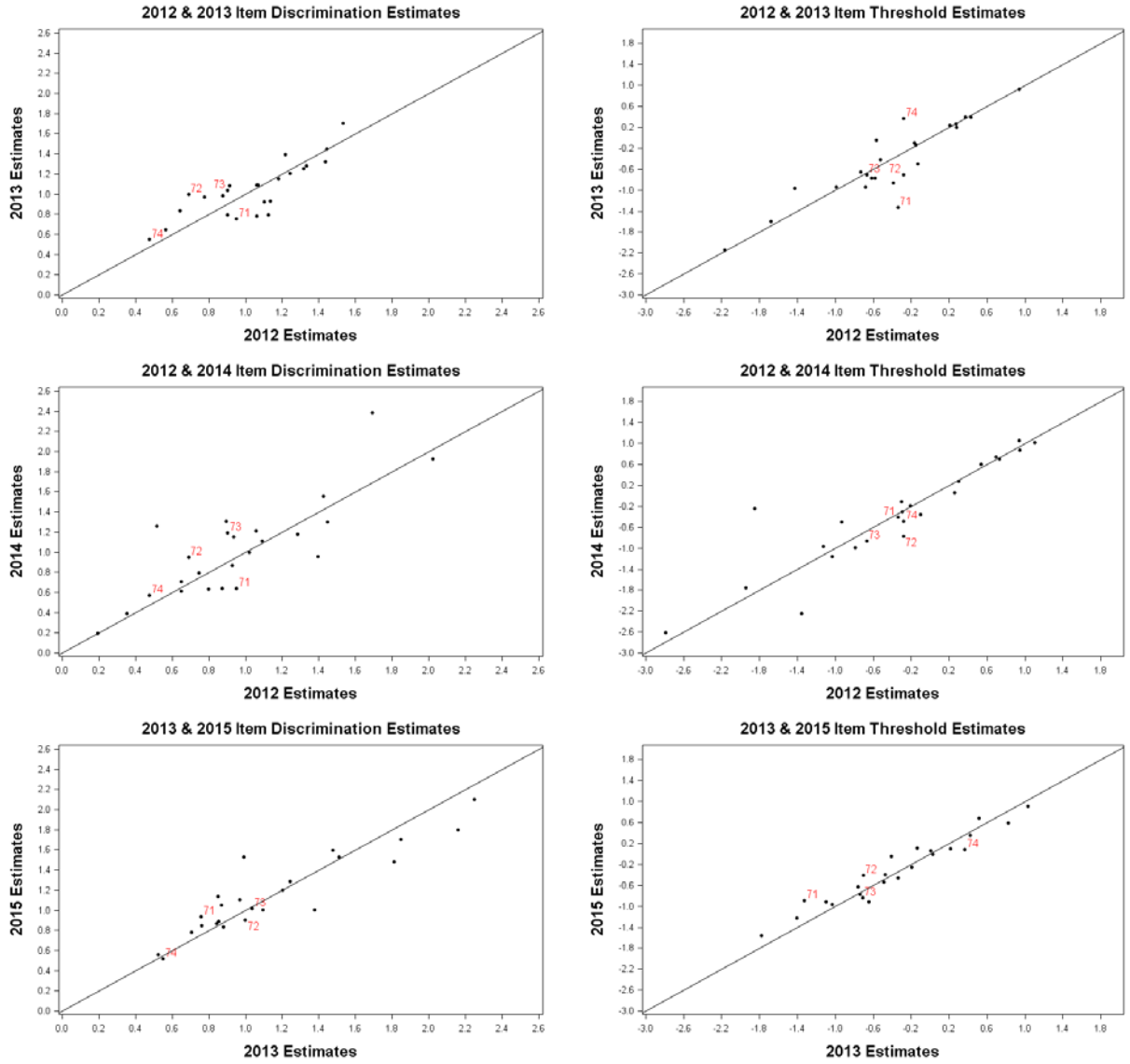

Figure 5. Visual inspection of old versus new form item parameter estimates for Italian Language using the original CINEG data.

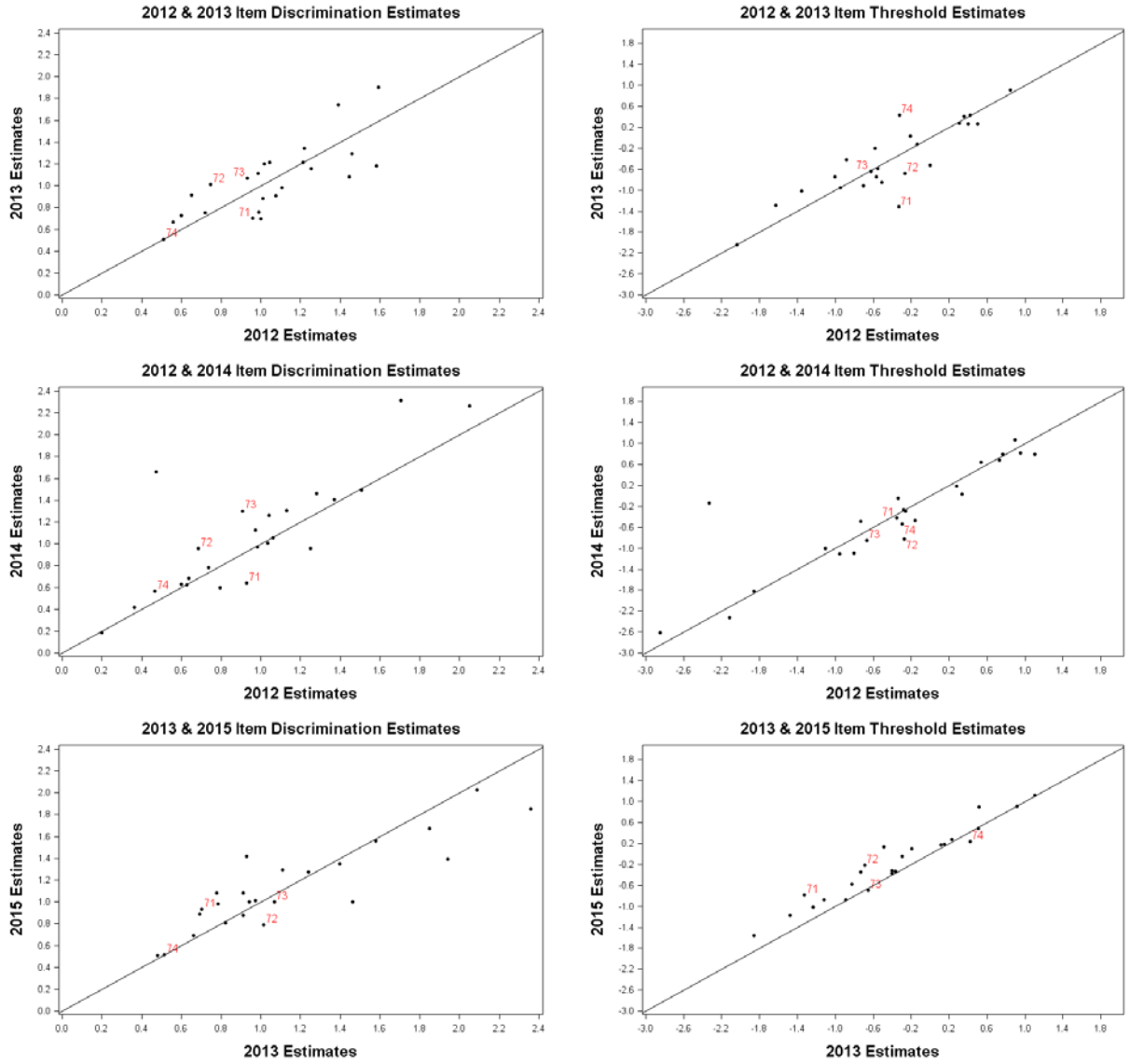

Figure 6. Visual inspection of old versus new form item parameter estimates for Italian Language using the RG datasets.

Chapter 5: Classification Consistency and Accuracy for Mixed-Format Tests

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

Abstract

This study explores classification consistency and accuracy for mixed-format tests using real data from various AP examinations. In particular, the primary purpose of the study is to compare results for various estimation procedures, based on either classical or IRT assumptions, that can be used for mixed-format data. In addition, classification indices are compared in relation to various factors: (a) test reliability, (b) test length, (c) cut score location (d) multidimensionality, and (d) score distributions. The results are presented for three selected classification indices including agreement index *P*, kappa coefficient, and classification accuracy index, gamma. This paper also proposes an algorithm that can be used for computing the conditional composite score distributions with non-integer weights. Finally, this study examines model-based score distributions for various estimation procedures, as an attempt to address an issue related to structural bumpiness in score distributions resulting from the use of non-integer section weights.

Classification Consistency and Accuracy for Mixed-Format Tests

The primary purpose of Advanced Placement (AP) examinations (The College Board, 2016) is to evaluate whether each examinee has achieved a certain level of performance and to award college credits to students who earn sufficiently high scores. The interpretation of the scores is made based on the location of the examinee's test score relative to cut scores. Colleges and other educational institutions use AP test scores to make classification decisions. Since the consequences of such decisions are critical to examinees, it is important to ensure that such classifications are reliable and accurate. The terms classification consistency and accuracy are often used to refer to reliability and validity in the context of classification. Classification consistency refers to the extent to which examinees are consistently classified into the same categories on two independent administrations of a test; while classification accuracy refers to the extent to which the actual classifications based on observed scores agree with the true classifications based on known true scores (Lee, 2010).

Classification consistency can be estimated directly if two alternate forms of a test can be given to the same group of examinees. However, repeated testing is impractical due to the constraints such as double testing time for examinees and expenses that go into the construction of two parallel forms. Instead, a number of procedures have been proposed for estimating classification consistency and accuracy indices based on data from a single test administration (Brennan & Wan, 2004; Breyer & Lewis, 1994; Lee, 2008, 2010; Livingston & Lewis, 1995; Peng & Subkoviak, 1980). However, little research has been done to study those procedures in the context of mixed-format tests, which often require taking potential item-format effects into account. Moreover, there exist a few procedures in the existing body of literature that are applicable to mixed-format tests, but they have not been researched extensively or applied to real mixed-format tests. Potential issues related to estimating classification consistency and accuracy for mixed-format tests motivate the present study.

Research Objectives

The current study is designed to explore classification consistency and accuracy for mixed-format tests using real data from various AP examinations. In particular, classification indices are compared in relation to various levels of test reliability, test length, section weights, multidimensionality, and score distributions. In addition, the current study compares results for various estimation procedures, based on either classical or IRT assumptions, that can be used for

mixed-format data. One vexing problem that arises with IRT procedures when applied to a mixed-format test is computing conditional distributions for composite scores, which often are formed using non-integer weights for scores on multiple-choice (MC) and free-response (FR) sections. This paper proposes an algorithm for computing the conditional composite score distributions with non-integer weights. Also, the impact of multidimensionality (i.e., item format effect) on classification estimates is investigated. Finally, this study addresses the issue of structural bumpiness in score distributions resulting from the use of non-integer section weights, and explores how each estimation procedure deals with structural bumpiness through an examination of a fitted score distribution.

Estimating Classification Consistency and Accuracy

Six estimation procedures applicable to mixed-format tests are employed in this study, including the normal approximation (Peng & Subkoviak, 1980), Livingston-Lewis (Livingstion & Lewis, 1995), compound multinomial (Lee, 2008), unidimensional IRT (Lee, 2010), simple-structure IRT (Knupp, 2009), and bi-factor IRT (LaFond, 2014) procedures. Depending on the psychometric models used, the first three procedures can be viewed as classical approaches, whereas the last three are IRT approaches. In this section, each procedure is discussed in terms of the underlying logic, assumptions, and implementation in the analysis.

Normal Approximation Approach

Peng and Subkoviak (1980) proposed the normal approximation approach (NM) for estimating classification consistency which is an approximation to a more complicated normal approximation method suggested by Huynh (1976). The NM method is based on a strong assumption that scores from parallel tests follow a bivariate normal distribution with a correlation equal to test reliability (r). Using the mean and standard deviation of the observed score distribution, μ and σ, the proportion of the bivariate normal distribution below and above the cut score can be easily computed. This proportion is interpreted as the proportion of examinees consistently classified into the same performance level on two (randomly) parallel forms of a test. Several methods have been proposed regarding the type of reliability estimate used for the correlation between two distributions. Peng and Subkoviak preferred the KR-21 reliability estimate, whereas Woodruff and Sawyer (1989) recommended using half-test reliability which is adjusted using the Spearman-Brown prophecy formula to obtain the reliability estimate for the full-length test. Although Peng and Subkoviak (1980) did not consider

classification accuracy, similar assumptions could be used for estimating classification accuracy indices. It is assumed that the true and observed scores follow a bivariate normal distribution with a correlation between the true and observed scores equal to the square root of reliability ($\sqrt{r}$).

Livingston-Lewis Procedure

The Livingston-Lewis procedure (LL) (Livingston & Lewis, 1995) was developed with the intent of creating a generally applicable method for any type of test scores, including scores from essay and performance assessments. This procedure is similar to Hanson and Brennan (1990) in that the true scores are assumed to take the form of either a two or four-parameter beta distribution and the errors are assumed to follow a binomial distribution. The key difference is that the effective test length plays an important role for LL in estimating the conditional distributions. The purpose of employing the concept of the effective test length is to find the number of discrete, dichotomously scored, locally independent, equally difficult test items so that total scores have the same precision as the actual scores. The effective test length is obtained using the following equation:

$$int(n) = \frac{(\mu - X_{min})(X_{max} - \mu) - r\sigma^2}{\sigma^2(1 - r)}, \quad (1)$$

where $int(n)$ represents a rounded integer effective test length; X_{min} and X_{max} are the lowest and highest possible scores, respectively; and μ, σ, and r are defined previously. As the next step, a new scale is set up by transforming the test score, X, ranging from 0 to $int(n)$. The transformed score, X', is expressed as

$$X' = int(n)\frac{X - X_{max}}{X_{max} - X_{min}}. \quad (2)$$

The beta-binomial procedures (Hanson & Brennan, 1990) are then applied to the transformed scores. More specifically, the conditional distribution given each true score level is computed based on the transformed scores. By integrating the conditional distributions over the entire true score distribution, the marginal observed score distribution is computed, which can be viewed as a hypothetical observed score distribution for the other form of the test. Using the actual observed score distribution adjusted based on the effective test length and the model-based observed score distribution, classification consistency indices are computed. The estimated true

score distribution and the adjusted actual observed score distribution are used for computing classification accuracy indices.

Compound Multinomial Procedure

The compound multinomial procedure (CM) was developed for estimating classification indices for tests with complex item scoring. A multinomial model is used for tests composed of a single item set, and a compound multinomial model is used for tests with different sets of items (Lee, 2005a). An item set is defined as a bundle of items that have the same number of score categories. It is important to note that even for a mixed-format test, the number of item sets can be more than two because it is often the case that the number of score points differs between FR items.

Suppose that a test is composed of S sets of items each containing n_s (s = 1, 2, …, S) items. For each set s, items are scored as one of k_s possible score points and Y_{sl} (l = 1, 2, …, k_s) denotes the random variable indicating the number of items scored z_{sl} (l = 1, 2, …, k_s). It follows that $\sum_{l=1}^{k_s} Y_{sl} = n_s$, the total number of items is $\sum_{s=1}^{S} n_s = n$, and the total score for set s is $X_s = \sum_{l=1}^{k_s} z_{sl} Y_{sl}$. Finally, the weighted composite score across all item sets can be defined as $T = \sum_{s=1}^{S} w_s X_s$, with w_s representing the weight associated with set s. Letting $\boldsymbol{\pi_s} = \{\pi_{s1}, \pi_{s2}, \dots, \pi_{sk_s}\}$ represent the proportion of items in the universe for item set s for which an examinee can earn scores of $z_{s1}, z_{s2}, \dots, z_{sk_s}$, the probability of the total score X_s for item set s is given by

$$\Pr(X_s = x_s | \boldsymbol{\pi_s}) = \sum_{z_{s1}y_{s1}+z_{s2}y_{s2}+\cdots+z_{sk_s}y_{sk_s}=x_s} \Pr(Y_{s1} = y_{s1}, Y_{s2} = y_{s2}, \dots, Y_{sk_s} = y_{sk_s} | \boldsymbol{\pi_s}), \quad (3)$$

where the random variables $Y_{s1}, Y_{s2}, \dots, Y_{sk_s}$ can be modeled by a multinomial distribution:

$$\Pr(Y_{s1} = y_{s1}, Y_{s2} = y_{s2}, \dots, Y_{sk_s} = y_{sk_s} | \boldsymbol{\pi_s}) = \frac{n_s}{y_{s1}!\, y_{s2}! \dots y_{sk_s}!} \pi_{s1}^{y_{s1}} \pi_{s2}^{y_{s2}} \dots \pi_{sk_s}^{y_{sk_s}}. \quad (4)$$

The assumption of uncorrelated errors over item sets leads to

$$\Pr(X_1 = x_1, X_2 = x_2, \dots, X_S = x_S | \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \dots, \boldsymbol{\pi_s}) = \prod_{s=1}^{S} Pr(X_s = x_s | \boldsymbol{\pi_s}). \quad (5)$$

Finally, the probability density function of the random variable for the weighted composite score, T, over all sets is given by

$$\Pr(T = t|\ \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s}) = \sum_{x_1,\ldots,x_S:\sum w_s x_s = t} \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_S = x_S|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s}). \quad (6)$$

Note that the summation is taken over all possible combinations of $w_s x_s$ that lead to the particular composite score *t*.

Suppose a test has *J* mutually exclusive performance categories with a set of observed cut scores, $c_1, c_2, \ldots, c_{K-1}$. The probability of being classified into category I_j for examinee *p* is

$$\Pr\big(T \in I_j|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s}\big) = \sum_{t=c_{(j-1)}}^{c_j-1} \Pr(T = t|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s}). \quad (7)$$

Then, the classification consistency index for each individual examinee can be expressed as

$$\Pr\big(T_1 \in I_j, T_2 \in I_j|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s}\big) = \left[\sum_{t=c_{(j-1)}}^{c_j-1} \Pr(T = t|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_s})\right]^{\mathbf{2}}. \quad (8)$$

The overall index of classification consistency for a group of people can be obtained by averaging the conditional (individual) estimates in Equation 8 over all examinees.

Estimating classification accuracy using the CM procedure depends on the use of the observed score as an estimate of an examinee's true score. Consequently, an examinee's true classification is equivalent to his/her actual classification made based on the observed score. Therefore, the conditional classification accuracy estimate for the examinee becomes equal to the sum of model-based probabilities of all observed score points that belong to the category the examinee is actually assigned to.

IRT Approach

Three IRT models that can be applied to mixed-format tests are considered in this study: unidimensional IRT, simple-structure multidimensional IRT, and bi-factor multidimensional IRT. Whereas unidimensional IRT does not consider the item format effects in the sense that it assumes a test only measures a single ability, the other two MIRT procedures take item format effects into account with the assumption of multiple distinct abilities associated with each item-format section.

Unidimensional IRT procedure (UIRT). Lee (2010) describes a procedure for estimating classification indices using UIRT models that can be used for either dichotomously or polytomously scored items. Let $\Pr(T = t) = \int_{-\infty}^{\infty} \Pr(T = t|\theta)\, g(\theta) d\theta$ denote the probability of earning composite score *t*, where $\Pr(T = t|\theta)$ is the conditional observed score distribution, and

$g(\theta)$ is the latent trait density. The probability of achieving composite score *t* for an examinee with an ability level of θ can be simply determined by multiplication of the response probabilities for each item, relying on the local independence assumption of IRT.

Once the conditional distribution for a given θ, $\Pr(T = t|\theta)$, is obtained, the probability of being classified into a particular performance category can be computed for each θ similar to Equation (7) replacing $\boldsymbol{\pi}'\boldsymbol{s}$ with θ. Given the conditional observed score distribution, the conditional consistency and accuracy indices are computed. Then, marginal indices are computed by integrating the conditional indices over $g(\theta)$.

Simple-Structure Multidimensional IRT procedure (SS-MIRT). When a test is composed of different sets of items measuring different, yet, correlated abilities, the data structure can be adequately modeled by the SS-MIRT model. Lee and Brossman (2012) attempted to apply SS-MIRT to equating with mixed-format tests. Knupp (2009) proposed an SS-MIRT procedure for estimating classification consistency and accuracy for composite scores.

In SS-MIRT, each section score is fit with a UIRT model, and the two abilities, θ_{MC} and θ_{FR}, are allowed to be correlated. The probability of having the composite score *t* under SS-MIRT for mixed-format tests can be expressed as:

$$\Pr(T = t|\,\theta_{MC}, \theta_{FR}) = \sum_{w_{MC}x_{MC}+w_{FR}x_{FR}=\,t} \Pr(X_{MC} = x_{MC}|\theta_{MC})\,\Pr(X_{FR} = x_{FR}|\theta_{FR}). \quad (9)$$

The conditional composite observed score distribution is computed for each pair of θ_{MC} and θ_{FR} using, typically, a bivariate normal distribution. Given the conditional composite score distribution, the conditional classification consistency and accuracy indices can be computed, which in turn are integrated (or summed) over the entire bivariate ability distribution to obtain the marginal indices.

Bi-Factor Multidimensional IRT procedure (BF-MIRT). The full-information item bi-factor model (Gibbons & Hedeker, 1992; Gibbons et al., 2007) has been one of the most popular models that can be applied to tests having a multidimensional data structure. In BF-MIRT framework, a mixed-format test is viewed as measuring two specific abilities associated with each item format, in addition to a general ability.

Recently, LaFond (2012) suggested a procedure for estimating classification consistency and accuracy using the bifactor and testlet response theory models for a testlet-based test. Although his study was conducted based on testlet-based tests, the BF-MIRT procedure can also

be implemented in the context of a mixed-format test. The probability of the composite score, *T*, can be expressed as

$$\Pr(T = t|\,\theta_g, \theta_{MC}, \theta_{FR}) = \sum_{w_{MC}x_{MC}+w_{FR}x_{FR}=\,t} \Pr\left(X_{MC} = x_{MC}\middle|\theta_g, \theta_{MC}\right) \Pr\left(X_{\mathrm{FR}} = x_{\mathrm{FR}}\middle|\theta_g, \theta_{FR}\right), \quad (10)$$

where θ_g represents the general ability and θ_{MC} and θ_{FR} indicate format-specific abilities. Note that each section score distribution is a function of a general ability and an item-format-specific ability.

Review of Relevant Literature

Several previous studies have explored various factors that could influence the estimation of classification indices. The factors that have been investigated in the literature include: (a) test length, (b) reliability, (c) observed score distribution, (d) dimensionality (e.g., testlet effects, construct equivalence, etc.), (e) cut score location, (f) number of classification categories, (g) score transformation, and (h) IRT model fit.

Regarding test length, it has been reported that as test length increases, classification consistency and accuracy increases and their estimates become more accurate and stable regardless of estimation procedures (Deng, 2011; Li, 2006; Wan, Brennan, & Lee, 2007). Wan et al. (2007) suggested that test length might serve as a mediating factor for test reliability. In other words, test length does not directly affect the classification indices but rather indirectly influences by changing test reliability. The effect of reliability has not been thoroughly investigated, although previous literature suggests that higher reliability leads to higher classification indices (Deng, 2011). The formal relationship between classical test reliability and classification consistency has not yet been well established.

Depending upon the psychometric models used, the observed score distribution can also affect the classification estimates. For instance, for the NM procedure, when the observed scores do not follow the normal distribution, it has been found that a considerable degree of bias can be introduced into kappa coefficients (Peng & Subkoviak, 1980; Woodruff & Sawyer, 1989). Similarly, when the true score distribution is bimodal, the LL procedure has been known to provide inaccurate estimates (Livingston & Lewis, 1995). In terms of dimensionality, previous findings are somewhat inconsistent. Wan et al. (2007) found that the impact of construct equivalence between two item formats was negligible, especially for *P* estimates – a similar amount of error was produced regardless of the degree of dimensionality. In contrast, LaFond

(2014) reported that the degree of the testlet effect significantly influenced the performances of UIRT and BF-MIRT. When the testlet effect was low, the UIRT method performed better, whereas under the high testlet effect condition, the BF-MIRT method outperformed the UIRT method. The reason for this discrepancy might be due in part to the fact that the two studies employed different psychometric models. The former only investigated classical models (i.e., non-IRT models), for which none of the estimation procedures considered the dimensionality of a test, whereas the latter compared the performances between UIRT and MIRT models, for which only one took into account potential dimensionality, whereas the other two did not.

It has been seen that the position of the cut score has a substantial effect on the estimation of classification consistency and accuracy, although the impact does not straightforwardly depend on the statistic estimated and the estimation procedure used. In general, a cut score near the mean or median (i.e., the point where relatively a large number of examinees are located) leads to lower *P* estimates but higher kappa estimates (Huynh, 1976; Knupp, 2009; Lee, 2008; Wan et al., 2007).

As the number of classification categories increases, the classification consistency and accuracy estimates tend to be lower by allowing more classification errors (Berk, 1980; Deng, 2011; Feldt & Brennan, 1989; LaFond, 2014; Wan, 2006; Wan et al., 2007). In terms of score transformation, when the transformation from raw to scale scores was based on a one-to-one relationship, the impact of scale transformation was minimal (Deng, 2011; Knupp, 2009; Wan et al., 2007). Regarding the effects of model fit, there have been mixed findings in the literature. Deng (2011) observed a severe impact of model misfit for IRT procedures, whereas Lee (2010) found that the effects of model fit for UIRT procedures were not clear.

Method

Data

The tests used for this study were seven AP examinations administered in 2014, including United States History, Biology, Chemistry, English Language and Culture, French Language and Culture, German Language and Culture, and Spanish Language and Culture. All these examinations are mixed-format tests composed of both MC and FR items. The number of MC and FR items and section weights vary by examination, and are summarized in Table 1. Originally, US History had three FR items with two different section weights (4.5 for two items and 2.75 for the remaining one). Since the CM procedure cannot be applied to data with an item

set with a single item, some arbitrary manipulation of weights was made and weights of 3.33 were used for all three items. Sample size varied, ranging from 4,283 to 17,969. Descriptive statistics for the datasets are reported in Table 2.

As an attempt to assess the dimensionality of the AP exams due to item format, the disattenuated correlations between MC and FR section scores ($\hat{\rho}_{\theta_{MC}\theta_{FR}}$) were computed. Small values of disattenuated correlation indicate the two sections measure different abilities, implying that the data are multidimensional. In this study, disattenuated correlation ranged from .75 to .97, with the smallest value for English Language and Culture and the largest value for Chemistry. Disattenuated correlation values are presented in Table 3. As discussed in Chapter 1 of this monograph, although the research in this chapter was conducted using data from the AP Examinations, the data were manipulated in such a way that the research does not pertain directly to operational AP examinations.

Classification Indices

Agreement indices *P* and Kappa were used in this study as classification consistency indices. The agreement index *P* is defined as the proportion of examinees consistently classified on two parallel forms of a test (Hambleton & Novick, 1973). Another well-known index, kappa coefficient *k*, (Cohen, 1960), corrects for chance agreement.

The gamma index (Lee, 2010) was used as a classification accuracy index. The gamma index represents the proportion of correctly categorized examinees and can be computed by comparing actual classifications based on the observed score distribution and true classifications based on examinees' estimated true scores. In this sense, it is similar to the *P* index, but it is based on a bivariate frequency distribution of the observed and true scores. Note that in order to avoid unnecessary repetition, a detailed description of computation of the three indices is omitted here. Readers might refer to Chapter 6 for more detailed information about these indices.

Classification Categories and Cut Scores

In scoring AP exams, each student is classified into one of five categories based on four cut scores. The cut scores used in this study were similar to the operational cut scores, but not necessarily the same. This study considered both binary classifications (i.e., pass or fail) and multi-level classifications. For multi-level classification, all four cut scores were applied simultaneously. Cut score information is presented in Table 3.

Estimation Procedures

Six estimation procedures, including both classical and IRT approaches, were applied and compared. This section presents how each estimation procedure was implemented.

Normal approximation procedure. Estimating classification indices using the NM procedure requires use of a reliability coefficient. In the context of classification, the main interest is to find each examinee's performance category with respect to the pre-specified cut scores. Therefore, reliability estimates that involve absolute error variance would be a reasonable choice. In this sense, reliability coefficients estimated from the CM model (Lee, 2005a) or Φ (phi) coefficient in generalizability theory (Brennan, 2001) could be possible candidates, and the former was used in this study. Reliability estimates for each examination are summarized in Table 3. Implementation of the NM procedure was undertaken using R.

Livingston-Lewis procedure. To implement the LL procedure, four pieces of information are required as input: (a) the observed score distribution, (b) a reliability coefficient of the scores, (c) the maximum and minimum possible scores on the test, and (d) cut scores. As with the NM procedure, reliability coefficients could be obtained using the CM model.

Analysis was conducted using the BB-CLASS program (Brennan, 2004). Traditionally, classification estimates using the LL method are computed by comparing examinees' scores on the actual form and a hypothetical form (i.e., the actual observed score distribution and the hypothetical observed score distribution predicted from the model). BB-CLASS also provides results based on two hypothetical forms. The former approach was employed in this study because it was the approach that was suggested in the original literature (Livingston & Lewis, 1995).

Compound multinomial procedure. The computer program MULT-CLASS (Lee, 2005b) was used to execute the CM procedure. Bias-correction (Brennan & Lee, 2006) was used to reduce bias resulting from the use of observed proportion correct score(s) as an estimate of each person's true proportion correct score(s). Wan et al. (2007) extended the bias-correction method for use with mixed-format data. They found that bias correction method could substantially reduce bias in classification consistency estimates and provided better results (Brennan & Lee, 2006; Wan et al., 2007).

IRT procedures. Using IRT procedures for estimating classification consistency and accuracy requires item and person parameter estimates. The computer program flexMIRT (Cai, 2012) was used to obtain parameter estimates for UIRT, SS-MIRT, and BF-MIRT models.

The computer program mIRT-CLASS (Lee, 2014) was used to estimate classification consistency and accuracy. For these analyses, both the D and P methods (Lee, 2010) can be used to compute marginal results. The essential part of estimating classification indices using IRT procedures is to approximate the integral associated with the θ distribution. The D method, which is also called the distributional approach, uses estimated quadrature points and weights. The P method, or individual approach, employs individual θ estimates. With the P method, the classification indices for each examinee are calculated first and then averaged over all the persons in the data. Previous literature indicates that both approaches tend to produce comparable results (Lee, 2010), so only the D method was chosen for this study. One drawback of the D method is that it can be computationally intensive for multidimensional models because, as the number of dimensions increases, the number of possible combinations of quadrature points required for calculation increases dramatically. However, since this study only deals with the mixed-format data structure, the maximum number of dimensions was three for BF-MIRT, which led to successful computations without excessive work.

Non-integer Section Weights for IRT and CM

A fundamental step for estimating classification consistency and accuracy using either CM or IRT procedures is to obtain the conditional distribution. For a mixed-format test, the conditional distribution is obtained for composite scores. Typically, the Lord-Wingersky recursive formula (1984) or its extended version (Hanson, 1994) could be used for computing the composite-score conditional distribution. However, the Lord-Wingersky algorithm is not appropriate when the scoring units are a non-integer real numbers and different weights are applied to the items. To address this issue, a generalized algorithm for real-number item scores was recently proposed (Kim, 2013). For all the tests used in this study, the composite score is defined as the weighted sum of two section scores and each section is associated with a non-integer weight. The non-integer composite scores are then rounded to obtain rounded integer composite scores, which are the score scale of interest for analyses. The procedure presented as follows provides an efficient way of obtaining the conditional integer composite score

distribution. The UIRT procedure is used for illustration, but similar steps could be used for the CM method. The steps are:

1. Compute the conditional distribution for each section separately without considering section weights.
2. Calculate the probability of getting each composite score *t*, using

$$\Pr(T = t|\theta) = \sum_{t=\boldsymbol{int}(w_{MC}X_{MC}+w_{FR}X_{FR})} f_{MC}(x_{MC}|\theta) \cdot f_{FR}(x_{FR}|\theta), \quad (11)$$

   where the summation is taken over all possible pairs of $w_{MC}X_{MC}$ and $w_{FR}X_{FR}$ that produce a particular integer score *t* after rounding weighted composite scores to the nearest integer value.
3. Obtain the marginal distribution of weighted integer composite score *T* given by

$$f(T = t) = \int_{-\infty}^{\infty} f(T = t|\theta)\, g(\theta) d\theta, \quad (12)$$

   by integrating the conditional composite score distributions over the θ distribution.

The application of this algorithm to MIRT procedures is straightforward. For the SS-MIRT procedure, the conditional distribution is computed for each section using only one ability (i.e., θ_{MC} or θ_{FR}) in Step 1. Then, the probability of earning composite score *t* becomes

$$\Pr(T = t|\, \theta_{MC}, \theta_{FR}) = \sum_{t=\boldsymbol{int}(w_{MC}x_{MC}+w_{FR}x_{FR})} f_{MC}(x_{MC}|\theta_{MC})\, f_{FR}(x_{FR}|\theta_{FR}). \quad (13)$$

Similarly, in the BF-MIRT procedure, the conditional distribution for each section is computed using the general ability, θ_g, and one format-specific ability, either θ_{MC} or θ_{FR} in Step 1. The probability of the composite score *t* can be obtained in Step 2, as follows:

$$\Pr(T = t|\, \theta_g, \theta_{MC}, \theta_{FR}) = \sum_{t=\boldsymbol{int}(w_{\mathrm{MC}}x_{\mathrm{MC}}+w_{\mathrm{FR}}x_{\mathrm{FR}})} f_{MC}\left(x_{MC}|\theta_g, \theta_{MC}\right) f_{FR}\left(x_{\mathrm{FR}}|\theta_g, \theta_{FR}\right). \quad (14)$$

The subsequent steps are identical to those described previously.

Results

Results of this study are summarized in this section. First, the comparison of estimation procedures is presented. The, the impacts of studied factors including test reliability, test length, and cut score location are provided. Also, the effects of multidimensionality due to item format

effects are discussed. Last, the model-based observed score distribution for each method is examined in the presence of structural bumpiness in score distributions.

Comparison of Estimation Procedures

The classification consistency and accuracy estimates based on six estimation procedures are summarized in Table 4. To facilitate the comparisons across the estimation procedures, a visual representation of Table 4 is given in Figure 1. The first plot in Figure 1 displays the results for classification agreement index *P* for all examinations using six procedures when all cut scores were applied simultaneously. As shown in this figure, all procedures provide similar results, with a maximum difference of approximately .05 in Biology and English examinations.

The patterns across procedures are fairly consistent for all examinations. CM produces the largest estimates whereas LL has the smallest values, with the exception of Biology. This is somewhat unexpected because previous studies have consistently found that the IRT procedure tended to have larger values than the CM procedure (Lee, 2010; Knupp, 2009). Lee (2010) provides one possible explanation for this finding. The assumptions of binomial and compound multinomial models associated with the LL and CM procedures imply that the hypothetical test forms are randomly parallel. On the other hand, under the IRT procedures, test forms are assumed to have exactly the same item parameters, indicating that the test forms are strictly parallel. As a result, the procedure associated with the stronger assumption about the errors (i.e., IRT) should produce larger classification estimates. However, this explanation is not justified by the results found in this study. The CM procedure tends to produce larger estimates than three IRT procedures across all examinations, except for Biology. One possible explanation for this unexpected trend is the potential bias for the CM procedure. The CM procedure has been reported in the literature to have large positive bias for multi-level classifications (Wan et al., 2007). Subkoviak procedure (Subkoviak, 1976), which is the CM version for dichotomous items, has also been found to have a similar tendency in previous studies (Algina & Noe, 1978; Subkoviak, 1978).

Another noticeable pattern in the plot for agreement index *P* is that both MIRT procedures have nearly identical results, whereas UIRT has somewhat different values from the MIRT procedures. This will be discussed in greater detail in the following section.

The results for the kappa can be seen in the middle of Figure 1. The tendencies for kappa largely agree with those for agreement index *P*, whereas the absolute values of kappa estimates

tend to be lower than *P* estimates. This makes sense, considering that kappa is corrected for chance agreement. Specifically, unless *P* is 1 (completely identical classifications on two forms), *k* is always smaller than *P* because the subtraction of chance agreement is performed on both the numerator and the denominator. This finding is in line with the previous studies (Wan, 2006; Wan et al., 2007).

The results for the classification accuracy estimates are depicted at the bottom of Figure 1. The estimates generally tend to have larger values than consistency estimates because classification consistency includes errors from two test administrations; on the other hand, classification accuracy only involves errors from one test form. For accuracy estimates, the LL procedure produces noticeably smaller values than other procedures. The NM procedure performs similarly to the IRT procedures for accuracy estimates, but produces more comparable consistency estimate results to the LL procedure.

Studied Factors

Several factors that might affect classification estimates are explored including test reliability, test length, and cut score location.

Test reliability. Figure 1 allows for a visual inspection of the effects of test reliability on classification indices. In this figure, the examinations are ordered from high to low by their reliability estimates, with German having the largest reliability estimate and Spanish the smallest. As expected, as the reliability decreases, classification estimates tend to decrease. The only exception to this pattern occurs between English and Spanish. The formal relationship between reliability and classification consistency/accuracy has not been much discussed in the literature yet, but the results show that there is at least some degree of relationship between the two.

Test length. Test length (i.e., score range), is summarized in Table 2. US history has the widest score range, while Chemistry has the narrowest score range. By looking at the multi-level classification results depicted in Figure 1, these two exams did not have either the largest or smallest estimates among the seven exams. In other words, the results of classification estimates across different examinations show that test length does not play a critical role in explaining the behavior of these estimates. Since the range of composite scores is created based on the weighted sum of two section scores, the section weights could have unintentionally distorted the relationship between classification estimates and test length. Thus, the number of raw-score

points for each exam is also given in Table 2. Even with the number of raw-score points, a clear pattern cannot be seen. This differs somewhat from the previous finding that, as test length becomes longer, classification consistency and accuracy gets higher (Deng, 2011; Li, 2006; Wan et al., 2007). However, it is important to note that this study was conducted with real data, so other potential confounding factors could not be controlled systematically. Therefore, no definite conclusions are made here.

Cut score location. The impact of cut score position can be seen in Figures 2 through 4. In each figure, results pertaining to the first cut (cut1) are presented in the top left plot, cut2 in the top right plot, cut3 in the bottom left plot, and cut4 in the bottom right plot. Previous studies (Deng, 2011; Huynh, 1976; Wan et al., 2007) pointed out that as the cut score moves away from the mean, *P* increases while kappa decreases. Thus, the effect of cut score location needs to be considered in conjunction with an examination of a score distribution. The observed composite score distributions for the seven examinations can be found in Figure 5. It is clear from Figure 5 that the observed score distributions for Spanish and German are highly negatively skewed, having their modes near the upper end of the distributions. The distributions for US History and English appear fairly symmetric and close to a normal distribution. The positively skewed distribution was observed only for Chemistry.

It seems from Figure 2 that the cut score location has a huge impact on the agreement index *P*. Spanish and German tend to have relatively large *P* estimates at cut1 and cut2 where few examinees are located, while having small values at cut3 and cut4 where frequencies are relatively high. A reverse relationship can be seen with Chemistry that shows a positively skewed distribution. This finding is in line with the literature (Deng, 2011; Huynh, 1976; Wan et al., 2007). Another finding is that the NM procedure tends to have larger estimates for cut1 and cut4 but smaller estimates for cut 2 and cut3, compared to the other procedures. In addition, for cut2, the largest estimates are consistently associated with the CM procedure across all examinations, whereas in the other cut score positions the differences among the methods are small. A possible explanation for this pattern can be found in a previous study (Wan et al., 2007). In Wan et al. (2007)'s simulation study, the CM procedure produced substantially larger bias for the 65% cut score (i.e., the middle of a score distribution), but it produced quite small bias for the 50% and 80% cut scores (i.e., both ends of a distribution).

The results for kappa and gamma estimates can be seen in Figures 3 and 4, respectively. The overall trends, particularly the results for the gamma estimates, seem very similar to the results for *P*. For the results of kappa, more variability in the estimates across the estimation procedures is found for cut1. Interestingly, a more clear positive relationship is observed for all cut scores between test reliability and kappa than between reliability and *P*. This pattern is in alignment with the finding in Wan et al. (2007). They found that kappa was affected more by the magnitude of reliability estimates than *P*.

Effects of Dimensionality (Item-Format Effects)

When a test measures more than one latent traits, it would be more reasonable to consider all of the constructs that are designed to be measured in estimating classification consistency and accuracy. In this study, disattenuated correlations between the MC and FR section scores were computed in order to assess the dimensionality due to item-format effects. The effects of dimensionality can be seen in Figure 6. Note that examinations are ordered along the horizontal axis by the degree of dimensionality (i.e., English had the lowest disattenuated correlation, whereas Chemistry had the largest). As expected, the differences in *P* estimates between UIRT and MIRT become more noticeable as the data become more multidimensional. In general, UIRT tends to yield smaller estimates than MIRT.

Regarding dimensionality effects, Wan et al. (2007) investigated the performance of five non-IRT procedures for mixed-format tests. In their simulation study, they explored the effects of construct equivalence (i.e., dimensionality) and found no clear relationship between *P* values and the degrees of cross-format correlation. The current study noted similar findings as seen in Figure 6. It is apparent that *P* values do not vary systematically with respect to format effects.

Using real data, it is difficult to make conclusions about which method performs better than others. However, it is still worth noting that the two approaches (i.e., UIRT and MIRT) produce different results as data become more multidimensional. LaFond (2014), in his simulation study, found that BF-MIRT outperformed UIRT in estimating classification consistency and accuracy indices when there was a large testlet effect. Although his study was conducted in the context of testlet-based tests, it may be reasonable to expect that, in the case of a high degree of multidimensionality, the MIRT procedures would provide more accurate results than the UIRT procedure. In order to draw firm conclusions about dimensionality effects due to

the format difference, however, it is desired to conduct a simulation study with certain known criterion classification indices.

Structural Bumpiness in Observed-Score Distributions

One potential issue related to estimating classification indices for mix-format tests is the structural bumpiness occurring in score distributions. To explore this issue, composite score distributions for seven examinations are plotted in Figure 5. Notice that most of the distributions show some bumpiness, especially for US History and English Language. Given the very large sample size for each examination (except German Language), this bumpiness is not just sampling error, but it is most likely due to the fact that the composite score is a weighted sum of MC and FR section scores. As summarized in Table 1, the section weights range from 1.00 to 3.33, implying that applying a large section weight (e.g., 3.25 or 3.33) may lead to relatively high frequencies for some score points of the composite scores. The bumpiness in the composite score distributions resulting from its scoring nature is referred to as "structural bumpiness" in this study. The clearest examples among the seven datasets are US History and English Language. Both examinations have large weights for FR sections (i.e., 3.33 and 3.0556, respectively) and the structural bumpiness for the two exams occurs fairly regularly throughout the score distributions. Another possible explanation for structural bumpiness is rounding. For the purposes of score reporting and psychometric analyses, composite scores are rounded to integers. Rounding is likely to cause an erratic pattern of the score distribution.

When structural bumpiness occurs, the classification consistency and accuracy estimates are possibly affected because some model-based procedures such as LL and NM do not assume that model-based distributions have the potential to be bumpy. For example, the NM procedure assumes that observed scores from two forms follow a bivariate normal distribution. Likewise, the LL procedure employs a beta distribution for true scores and a binomial distribution for errors, which results in a smooth model-based observed score distribution. However, the IRT procedures can deal with structural bumpiness by incorporating the section weights systematically into the conditional probability computation process, as discussed previously. This observation is confirmed in Figure 7, which shows the model-based observed score distributions for four estimation procedures (NM, LL, UIRT, and SS-MIRT). Note that the results for CM and SS-MIRT were not included because CM does not use model-based observed score distributions and the SS-MIRT distribution was almost identical to the BF-MIRT

distribution. As explained, the model-based observed score distributions for the IRT procedures successfully follow the actual score distributions, whereas those for the NM and LL procedures are smooth. The score distribution for the LL procedure is based on the effective test length because this procedure employs the effective test length rather than the actual test length. Note, however, that this study employs real data analyses, and thus it is difficult to determine the extent to which structural bumpiness affects the accuracy of classification estimates for various estimation procedures. Reflecting the actual structural bumpiness in the model-based observed score distribution may not be a determining factor for the accuracy of classification consistency and accuracy estimates. This issue is fully examined with a simulation study in Chapter 6.

Discussion

When the main purpose of test administration is to identify an individual's status with respect to some established standard (e.g., cut scores, performance levels, etc.), a test score usually serves as an indicator as to whether an examinee possesses an acceptable level of skills or knowledge. Usually, such decisions are high-stakes, such as in program admission, certification, or licensure, resulting in significant consequences for both individuals and educational institutions. Therefore, it is necessary to assure that the test is reasonably reliable and valid in making such classification decisions. In recent years, many testing programs, including AP exams, are increasingly administering mixed-format tests. When a mixed-format test is used for classification purposes, classification consistency and accuracy need to be evaluated in consideration of the psychometric characteristics of the test format. This present study attempts to address this issue.

This study compared various classification estimation procedures that are applicable to mixed-format tests and investigated issues associated with mixed-format tests. More specifically, six existing estimation procedures were applied to seven AP examinations. This study used an efficient algorithm for computing conditional score distributions that can be used for non-integer section weights. Additionally, the impact of multidimensionality (i.e., item-format effect) was investigated. Finally, this study addressed a potential of structural bumpiness in observed score distributions for mixed-format tests, and explored the model-based score distribution for each estimation procedure to address the possibility of incorporating structural bumpiness in the estimation procedure. The ability of reflecting actual bumpiness for an estimation procedure,

however, may not necessarily lead to better estimation of classification indices. Due to the limitation of using real data, this question cannot be de answered straightforwardly in this study.

The results of this study led to the following conclusions. First, the results for all of the classical and IRT procedures show similar patterns across different exams. Second, the magnitudes of classification estimates are not substantially different across estimation procedures—in general, CM produces the largest consistency and accuracy estimates, whereas LL tends to have the smallest values. Third, non-integer section weights could effectively be treated with existing estimation procedures (e.g., CM and UIRT procedures). Fourth, classification indices are affected by (a) cut score location and (b) test reliability. Regarding the role of cut score location, it has been agreed upon from previous studies that the shape of the observed score distribution influences classification indices while interacting with the position of the cut score. Regarding the properties of examinations, there is some suggestion of a positive relationship between test reliability and classification consistency/accuracy, although the pattern is not entirely clear. Fifth, the relative behavior of the estimation procedures tends to be affected by the degree of multidimensionality in terms of latent correlation between the MC and FR sections. Consequently, as data become more multidimensional, UIRT and MIRT tend to produce different results, with UIRT yielding smaller estimates than MIRT.

As is true for any study using real data, the current study is limited in that there is no criterion classification consistency and accuracy. As a result, the relative performances of studied procedures could not be evaluated with respect to accuracy and precision. A more extensive systematic simulation study should be conducted to obtain a better understanding of the behavior of the existing procedures. Also, even though a close relationship between classical test reliability and classification consistency was observed, the theoretical and mathematical relationship has not yet been discussed in the literature, and thus is worth future research. Finally, in the future it would be worth exploring some other models such as full MIRT models (McKinley & Reckase, 1983; Reckase, 1985) for estimating classification consistency and accuracy for mixed-format tests.

References

Algina, J., & Noe, M. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement, 15*, 101-110.

Berk, R. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17*, 323-346.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy*, Version 1.1 (CASMA Research Report No. 9). Iowa City, IA: University of Iowa.

Brennan, R. L., & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City, IA: University of Iowa.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report 94-39). Princeton, NJ: Educational Testing Service.

Cai, L. (2012). *flexMIRT* (Version 1.88) [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Deng, N. (2011). *Evaluating IRT- and CTT-based methods of estimating classification consistency and accuracy indices from single administrations*. Unpublished dissertation. University of Massachusetts, Amherst, MA.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) *Educational measurement* (3th ed.). New York: American Council on Education and Macmillan.

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

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

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. *Journal of Educational Measurement, 10*, 159-170.

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

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Kim, S. (2013). Generalization of the Lord-Wingersky algorithm to computing the distribution of summed test scores based on real-number item scores. *Journal of Educational Measurement, 50*, 381-389.

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

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

Lee, W. (2005a). *A multinomial error model for tests with polytomous items*. (CASMA Research Report No. 10). Iowa City, IA: University of Iowa.

Lee, W. (2005b). *Manual for MULT-CLASS: For multinomial and compound multinomial classification consistency*. Iowa City, IA: University of Iowa.

Lee, W. (2008). *Classification consistency and accuracy under the compound multinomial model.* (CASMA Research Report No. 13). Iowa City, IA: University of Iowa.

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

Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33*, 374-390.

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

Lee, W. (2014). *mIRT-CLASS: A computer program for multidimensional item response theory classification consistency and accuracy* [computer software] (Version 1.0). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Li, S. (2006). *Evaluating the consistency and accuracy of proficiency classifications using item response theory.* Unpublished dissertation. University of Massachusetts, Amherst, MA.

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

McKinley, R. L. & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Rep. ONR 83-2). Iowa City IA: The American College Testing Program.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*, 359-368.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265-276.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement, 15,* 111-116.

The College Board. (2016, December 21). *AP Students – AP Courses and Exams for Students*. Retrieved from https://apstudent.collegeboard.org/home

Wan, L. (2006). *Estimating classification consistency for single-administration complex assessments using non-IRT procedures.* Unpublished doctoral dissertation. University of Iowa.

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments*. (CASMA Research Report No. 22). Iowa City, IA: University of Iowa.

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement, 13,* 33-43.

Table 1

Test Information and Sample Sizes

Exam	Section	# of Items	Score Points	Section Weights	*N*
German Language	MC	65	65	1.00	4,283
	FR	4	5,5,5,5	3.25	
Chemistry	MC	50	50	1.00	17,969
	FR	7	10,10,10,4,4,4,4	1.0869	
French Language	MC	65	65	1.0344	17,067
	FR	4	5,5,5,5	3.25	
US History	MC	80	80	1.125	17,239
	FR	3	9, 9, 9	3.33	
Biology	MC	58	58	1.00	9,911
	FR	8	10,10, 4,4,4,3,3,3	1.5 1.4285	
English Language	MC	55	55	1.2272	15,541
	FR	3	9,9,9	3.0556	
Spanish Language	MC	65	65	1.00	16,459
	FR	4	5,5,5,5	3.25	

Table 2

Descriptive Statistics for Data sets

Exam	Score Range	Score Points	Mean	Median	SD	Min	Max	Kurtosis	Skewness
German Language	0-130	85	90.911	93	25.502	12	130	2.382	-.390
Chemistry	0-100	96	44.443	43	19.598	4	98	2.162	.240
French Language	0-130	85	84.358	85	22.457	9	130	2.606	-.298
US History	0-180	107	85.479	86	26.169	10	168	2.548	-.022
Biology	0-120	109	67.200	69	21.232	8	117	2.352	-.238
English Language	0-150	82	80.254	81	20.248	10	143	2.865	-.204
Spanish Language	0-130	85	93.484	96	18.758	14	130	3.637	-.724

Table 3

Reliability and Cut Score Information

Exam	Reliability	Disattenuated Correlations	Cut Score (% of Examinees at)
German Language	.93797	.94	N/A(7.6), 52(17.7), 73(26.7), 95(23.3), 113(24.7)
Chemistry	.92818	.97	N/A(21.9), 27(25.4), 42(25.2), 58(16.8), 72(10.6)
French Language	.91807	.92	N/A(4.4), 44(15.9), 66(33.3), 88(26.6), 106(19.7)
US History	.91065	.89	N/A(16.9), 59(26.5), 82(21.8), 97(23.1), 118(11.6)
Biology	.88863	.96	N/A(6.4), 33(22.1), 55(32.5), 76(28.3), 94(10.7)
English Language	.82897	.75	N/A(9.9), 54(27.7), 75(30.1), 91(20.7), 105(11.6)
Spanish Language	.82014	.87	N/A(1.4), 43(7.8), 68(28.0), 90(36.5), 107(26.4)

Table 4

Classification Consistency and Accuracy Estimates for Seven Examinations using Six Estimation Procedures (Multi-level Classifications)

Index	Subject	NM	LL	CM	UIRT	SS-MIRT	BF-MIRT
P	German	.6923	.6805	.7183	.6883	.7015	.7020
	Chemistry	.6642	.6558	.6809	.6662	.6701	.6727
	French	.6769	.6660	.6965	.6697	.6826	.6850
	US History	.6185	.6138	.6524	.6131	.6340	.6342
	Biology	.6325	.6155	.6720	.6701	.6731	.6704
	English	.5305	.5182	.5687	.5174	.5521	.5533
	Spanish	.5990	.5663	.6180	.5947	.6086	.6161
Kappa	German	.5972	.5888	.6365	.5983	.6152	.6159
	Chemistry	.5634	.5610	.5925	.5730	.5779	.5812
	French	.5644	.5565	.5958	.5605	.5779	.5808
	US History	.5109	.5086	.5572	.5073	.5340	.5341
	Biology	.5003	.4873	.5602	.5562	.5601	.5586
	English	.3825	.3723	.4351	.3693	.4161	.4172
	Spanish	.4338	.3933	.4641	.4336	.4540	.4614
γ (Accuracy)	German	.7787	.7223	.7961	.7749	.7852	.7856
	Chemistry	.7569	.7408	.7685	.7553	.7622	.7635
	French	.7669	.7238	.7800	.7655	.7710	.7741
	US History	.7176	.7009	.7441	.7141	.7318	.7326
	Biology	.7319	.7038	.7634	.7632	.7642	.7636
	English	.6395	.6233	.6732	.6337	.6629	.6645
	Spanish	.7004	.6295	.7156	.7021	.7116	.7175

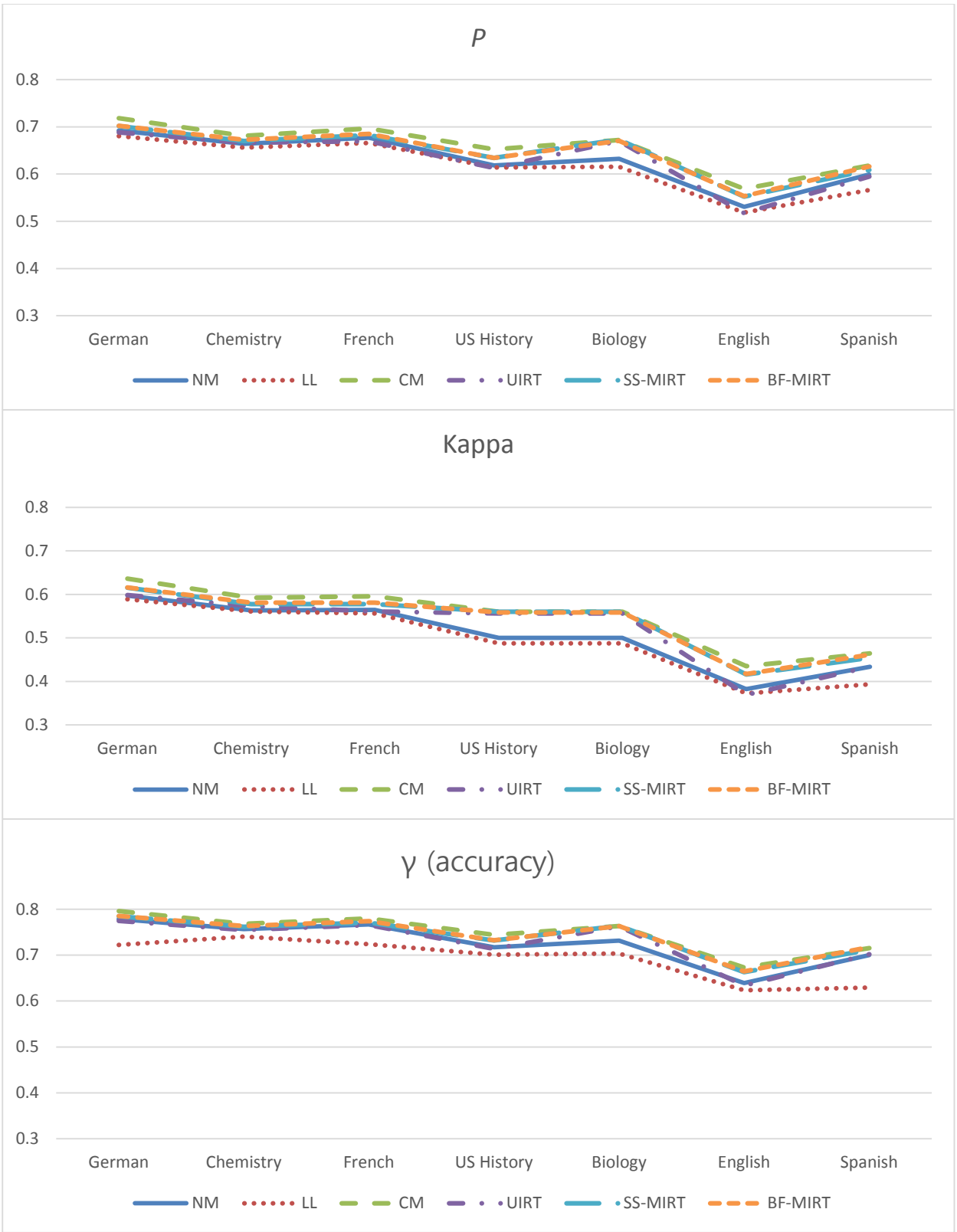

Figure 1. Classification consistency and accuracy estimates for multi-level classification.

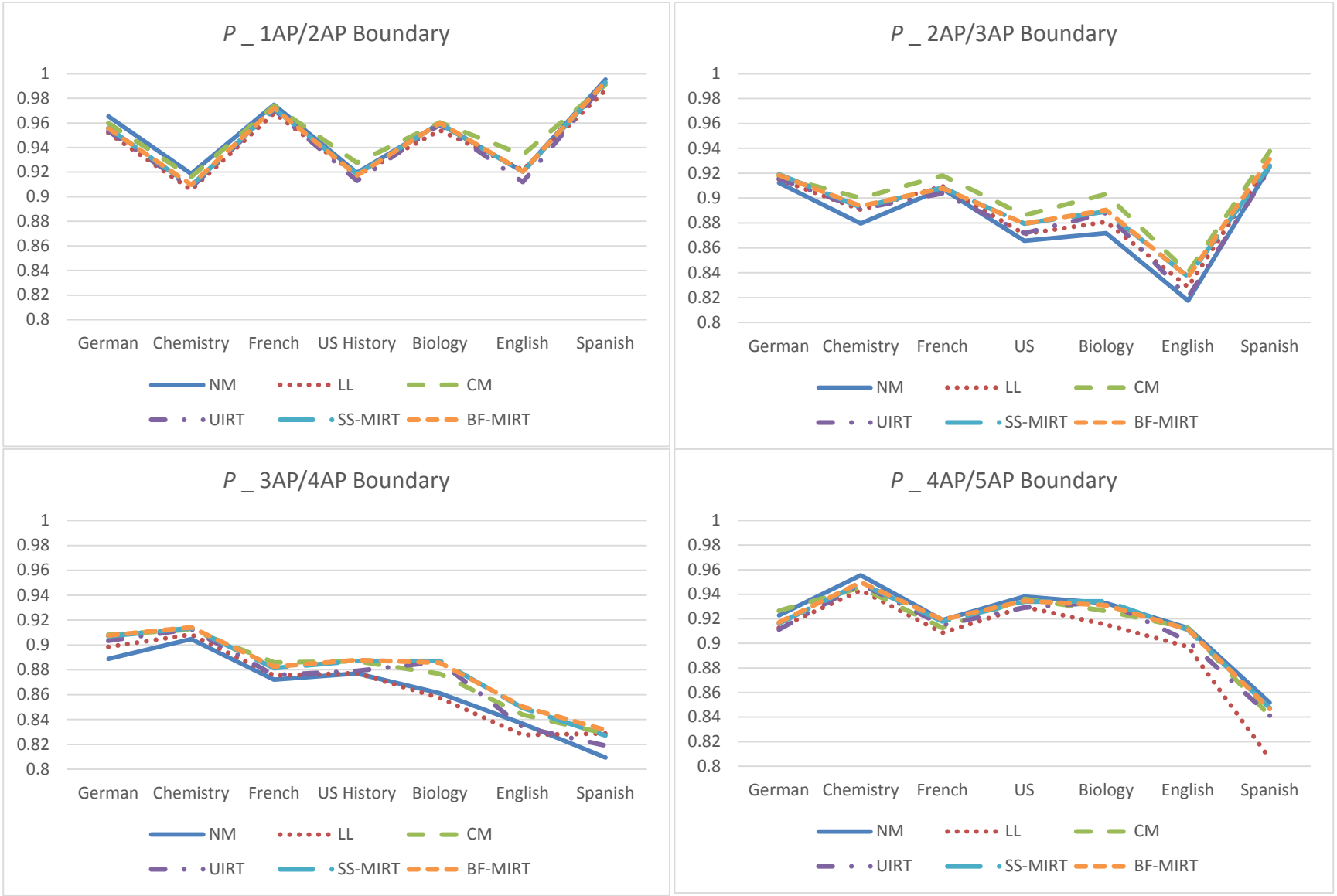

Figure 2. Agreement index *P* estimates for binary classifications.

Figure 3. Kappa estimates for binary classifications.

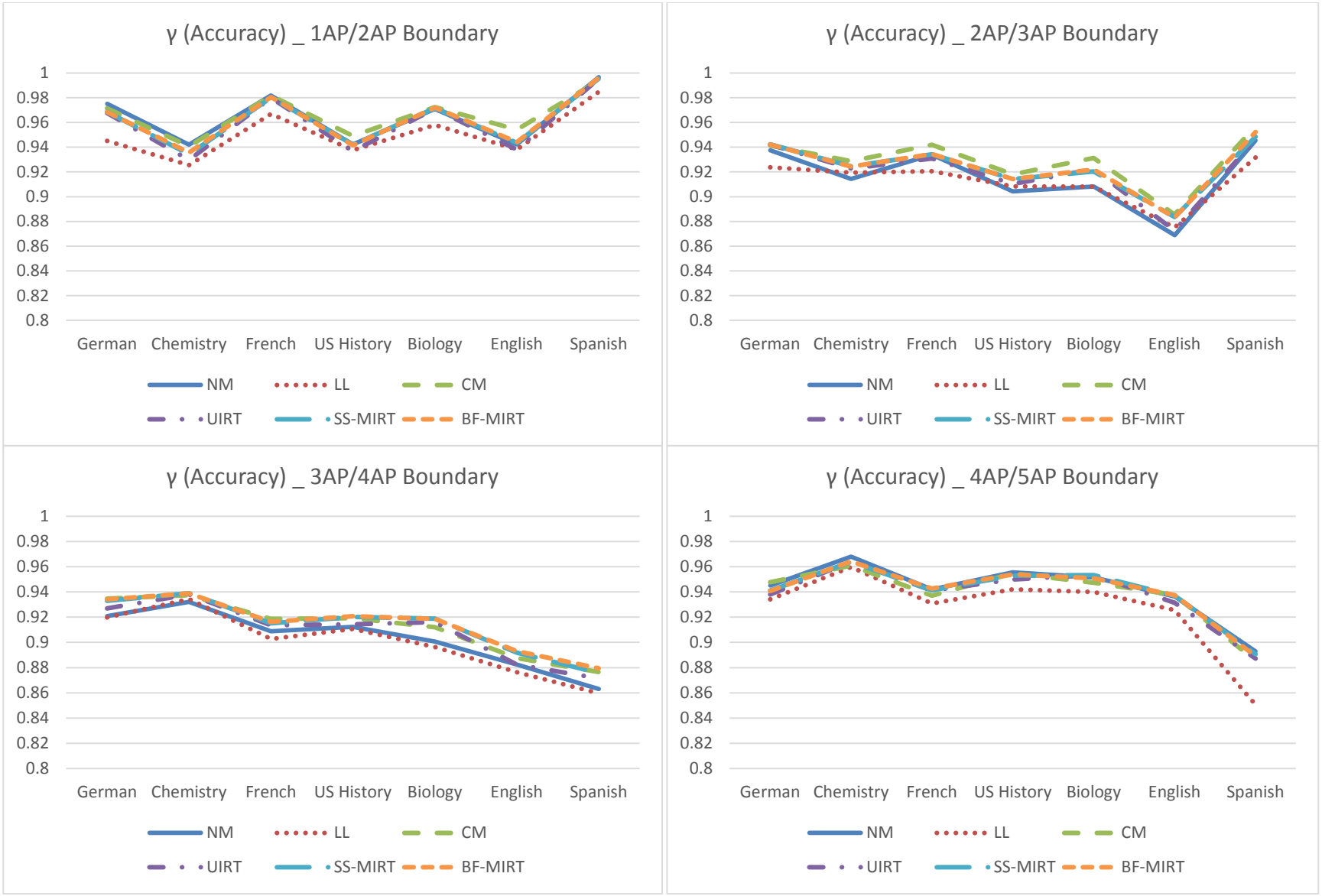

Figure 4. Gamma (accuracy) estimates for binary classifications.

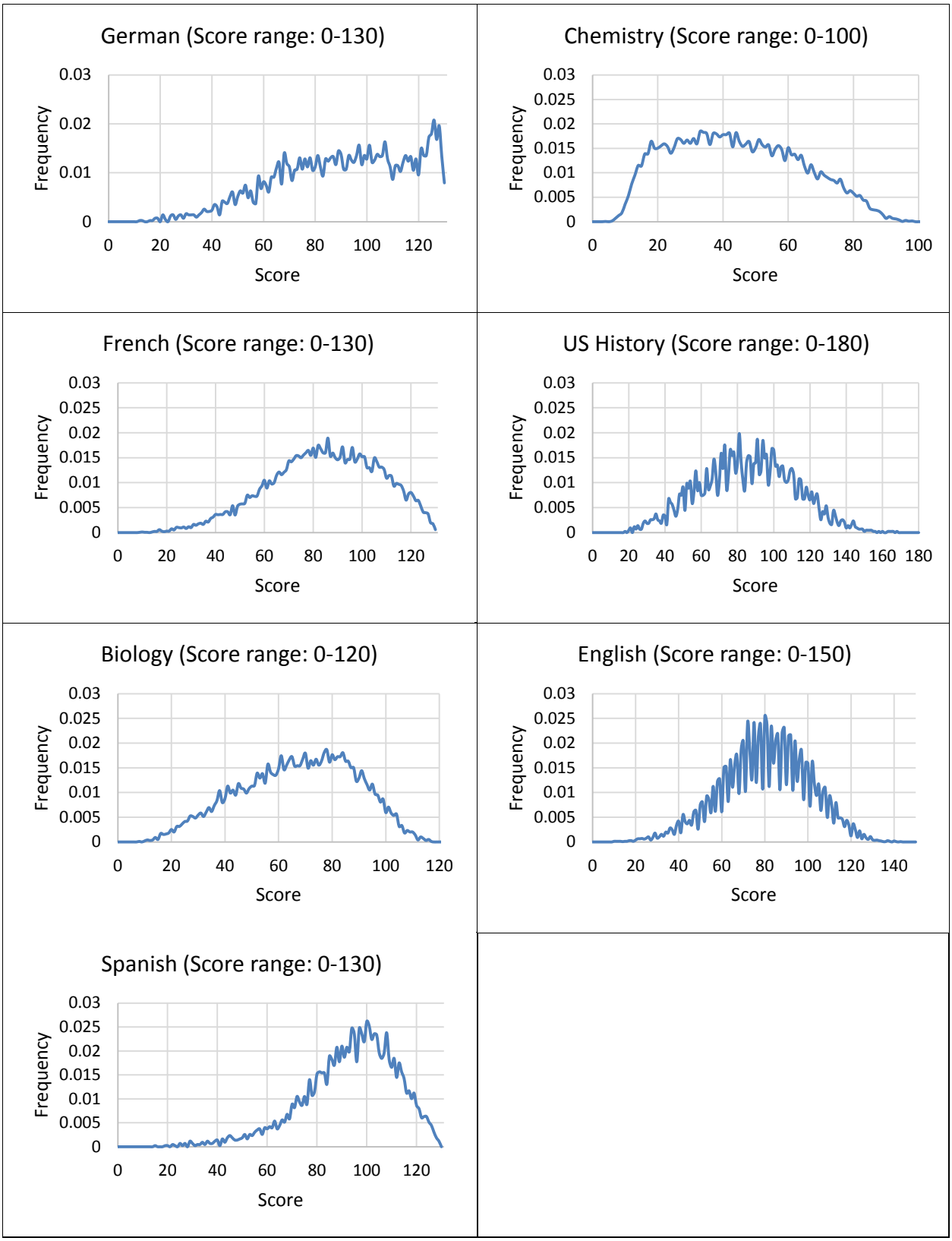

Figure 5. Observed composite score distributions for seven examinations.

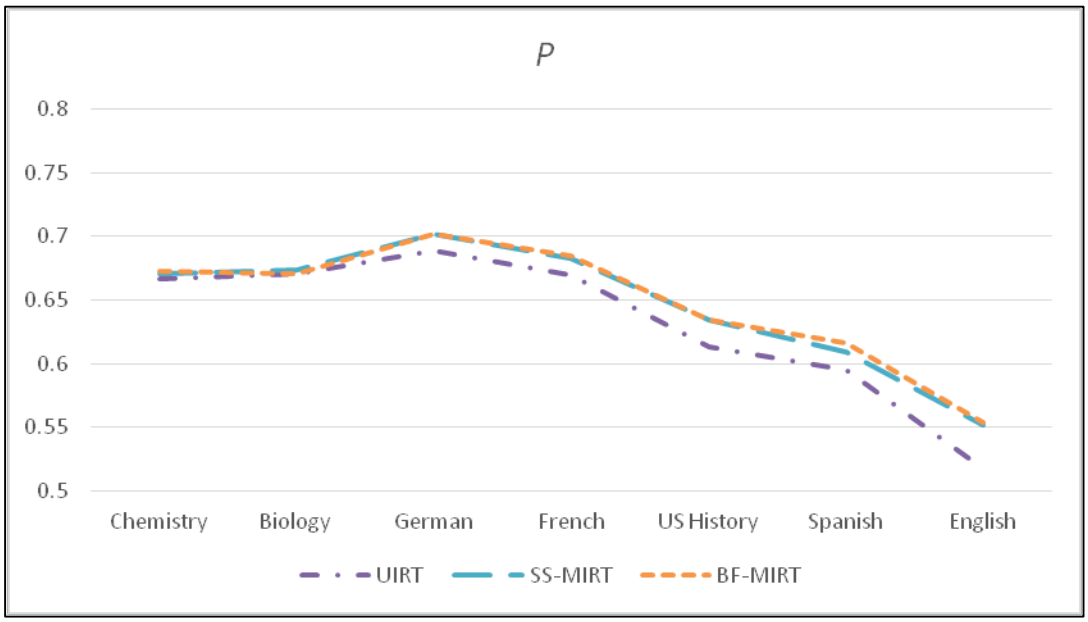

Figure 6. Effects of dimensionality (item-format effects) on *P* estimates among IRT procedures.

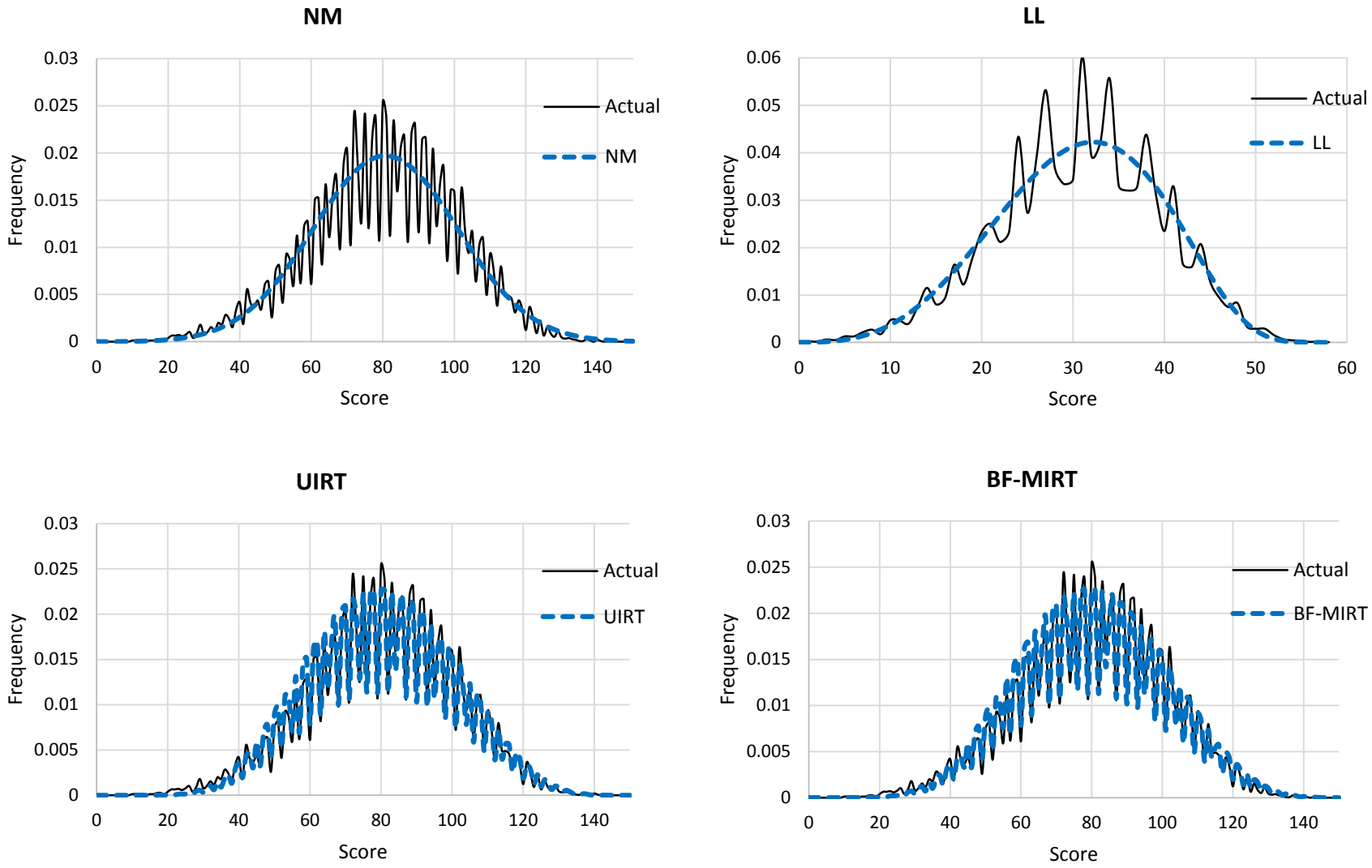

Figure 7. Model-fitted observed-score distributions.

Note. LL procedure is based on effective test length (58), not the actual test length.

Chapter 6: Classification Consistency and Accuracy with Atypical Score Distributions

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

Abstract

This study aims to evaluate the performance of three non-IRT procedures (i.e., normal approximation, Livingston and Lewis, and compound multinomial procedures) for estimating classification indices when the observed score distribution shows atypical patterns: (a) the observed score distribution is bimodal, (b) the observed score distribution has structural (i.e., systematic) bumpiness, or (c) the observed score distribution has some structural zeros (i.e., no frequencies). Under a bimodal distribution, the normal approximation procedure produced substantially large bias. For a distribution with structural bumpiness, the compound multinomial procedure tended to introduce remarkably large bias. Under a distribution with structural zeroes, the relative performance of selected estimation procedures depended on cut score location and the sample size conditions. Last, the largest standard error was consistently associated with the Livingston-Lewis procedure, while the smallest standard error tended to be observed for the normal approximation procedure across all studied distributions.

Classification Consistency and Accuracy with Atypical Score Distributions

Many large-scale testing programs, such as Advanced Placement (AP) Examinations (The College Board, 2016) and various types of licensure tests, administer a test designed to make individual or institutional decisions based on test score information. When the primary goal of a test is to make categorical decisions about examinees, it is crucial to assure that an examinee is properly categorized no matter which form is administered. The concept of classification consistency is often used to quantify the degree to which an examinee is consistently classified into the same performance level over alternate forms of a test. In addition to consistency, it is desired that examinees be accurately assigned into their true performance categories if their true scores are known. The psychometric concept of classification accuracy is often used to quantify how closely observed classifications based on examinees' observed scores coincide with the true classifications based on their true scores.

Some practical challenges in administering a test twice hinder direct computation of classification consistency between two test forms. Moreover, a practical issue in reporting classification accuracy is that examinees' true scores are never known and therefore, should be estimated. Therefore, various procedures that require only a single test administration have been developed for estimating classification consistency and accuracy indices. Huynh (1976) developed a procedure that can be applied to dichotomously scored data assuming a beta-binomial model, which was further extended by Hanson and Brennan (1990). Peng and Subkoviak (1980) extended Huynh (1976)'s work by assuming a bivariate normal distribution of scores from two independent test administrations with a KR-21 coefficient as a correlation. This procedure has been reported to be computationally amenable and produce reasonably accurate results (Wan, Brennan, & Lee, 2007). Lee (2005a) and Lee, Brennan, and Wan (2009) proposed a compound multinomial procedure that can be used for a mixed-format test. In this procedure, a multinomial error model is employed for a test containing a single item type (e.g., all items are scored 0-4); while a compound multinomial model is used for a test composed of a mixture of different item types (e.g., multiple-choice items scored 0/1 and essay items scored 0-4). Livingston and Lewis (1995) extended the beta-binomial procedures to make them applicable to polytomously scored data or other complex types of scores. The so called "effective test length" is computed to approximate the number of dichotomously scored items, which is then used to estimate the conditional distribution of scores on an alternate form.

With respect to the performance of procedures that are not based on item response theory (IRT), Wan et al. (2007) explored the behavior of five non-IRT procedures including the normal approximation procedure, the Breyer-Lewis procedure, the Livingston-Lewis procedure, a bootstrap procedure, and the compound multinomial procedure. Study conditions considered include test length, degrees of cross-format equivalence (i.e., dimensionality), cut score positions, and number of classification categories. This study is one of few studies that offer comprehensive comparison and evaluation of the performance of various non-IRT estimation procedures. However, there is still room for a closer examination of their behaviors under other important factors such as the shape of a score distribution, which is the main focus of this study.

Previous studies have noted that score distributions might be a potential factor affecting classification consistency and accuracy indices (Deng, 2011; Li, 2006). Specifically, Li (2006) explored the effects of distribution shape on classification consistency by varying the levels of skewedness and demonstrated that when test scores are negatively skewed, the difference among estimation procedures (i.e., three IRT procedures and Livingston & Lewis procedure) became larger. This implies that the score distribution could influence classification estimates. However, other than Li (2006)'s work, no studies have been conducted to examine the impact of various types of score distributions on classification consistency and accuracy estimates. Moreover, as noted by Livingston and Lewis (1995), their procedure may not perform accurately if the score distribution follows a bimodal distribution, which although has not yet been examined in the literature. This study aims to evaluate the performance of various estimation procedures under various "atypical" score distributions using three simulation studies.

Research Objectives

The primary objective of this study was to evaluate the performance of a few selected procedures for estimating classification indices when the observed score distribution shows atypical patterns, namely: (a) the observed score distribution is bimodal, (b) the observed score distribution has structural (i.e., systematic) bumpiness, or (c) the observed score distribution has some structural zeros (i.e., no frequencies). The three types of distributions considered in this study may not be seen very often in typical achievement tests, but they could occur in psychological tests and in achievement tests employing mixed item types and complex scoring. The relative performance of various procedures when applied to different types of observed score distributions is not currently known.

Estimation Procedures

Three estimation procedures were considered in this study: the normal approximation, Livingston-Lewis, and compound multinomial procedures. These three procedures were selected because none of them are based on IRT and they have been investigated most widely in previous studies. However, they do differ in that they are based on different psychometric models and require different assumptions about score distributions, which is mainly discussed in this section. Note that to avoid redundancy, specific steps for estimating classification indices are not provided in this chapter. Readers might refer to Chapter 5 for more details.

Normal Approximation Procedure

The Normal Approximation (NM) procedure (Peng & Subkoviak, 1980) is an extension of Huynh's work (1976) by approximating a bivariate beta-binomial model with a bivariate normal distribution. The beta-binomial model assumes a beta distribution for true scores and a binomial distribution for errors. The resultant observed score has the form of a negative hypergeometric distribution. This approach implicitly regards test forms as randomly parallel, treating items randomly selected from an undifferentiated universe of such items.

In the NM procedure, two alternate forms are considered to have an identical mean and standard deviation with reliability as a correlation between two forms. The form of the resultant bivariate normal distribution is highly dependent on the mean and standard deviation of the observed score distribution and, the distribution is always smooth without bumpiness. If the normal assumption does not hold for a score distribution, then the estimated classification indices might be inaccurate.

Livingston and Lewis Procedure

Livingston and Lewis (LL) procedure (Livingston & Lewis, 1995) is based mainly on the bivariate beta model proposed by Huynh (1976). In this procedure, instead of using the original dataset, a substitute test is created based on "effective test length." The effective test length is determined by a function of mean, maximum score, minimum score, reliability, and variance obtained from the original data. The concept of effective test length was introduced in order to incorporate various types of data, such as performance assessment data. The original form is transformed such that the hypothetical form is solely composed of dichotomously-scored items, having identical measurement precision with the original data. A four parameter beta binomial model (Hanson & Brennan, 1990) is then applied to the substitute test.

As one might easily anticipate, the model-based distribution is almost always smooth to take the form of a negative hypergeometric distribution without bumpiness. As already discussed, the LL procedure may not work adequately for a bimodal observed score distribution (Livingston & Lewis, 1995) because the beta distribution cannot fit a bimodal distribution well.

Compound Multinomial Procedure

To accommodate various types of data such as a mixture of dichotomously and polytomously scored items or a test involving polytomously-score items with different numbers of score categories, a multinomial model is adopted for a test with a single set of items and a compound multinomial model is employed for a test having multiple distinct sets of items (Lee, 2005a; Lee et al., 2009). Similar to a beta-binomial model, the compound multinomial (CM) procedure assumes that items are randomly drawn from a universe of undifferentiated items, which implies a randomly parallel assumption for test forms. Since the CM procedure does not require any specific distribution for true scores, there is no process for fitting the observed score distribution.

Classification Indices

In this study, two well-known indices of classification consistency are considered: agreement index *P (*Hambleton & Novick, 1973) and kappa coefficient (Cohen, 1960). Also, as a classification accuracy index, the gamma index (Lee, 2010) is used. A brief description of each of these indices is provided next.

Agreement Index *P*

The agreement index *P* is defined as the proportion of examinees classified into the same performance categories on two parallel forms of a test (Hambleton & Novick, 1973). It is denoted as

$$P = \sum_{j=1}^{J} P_{jj},$$

where *J* represents the number of performance categories and P_{jj} indicates the proportion of examinees assigned to the same j^{th} category on the two parallel forms.

Kappa Coefficient

Kappa coefficient (Cohen, 1980) was developed to correct for chance agreement. The correction factor, P_c, is presented as

$$P_c = \sum_{j=1}^{J} P_{j\cdot} P_{\cdot j},$$

where $P_{j\cdot}$ and $P_{\cdot j}$ indicate the marginal proportions of examinees falling into the j^{th} category in either of two forms. The kappa coefficient, k, is computed as

$$k = \frac{P - p_c}{1 - p_c},$$

where P is the agreement index.

Gamma Index

The gamma index proposed by Lee (2010) quantifies the proportion of examinees correctly classified into the "true" performance category, which is denoted by

$$\gamma = \sum_{j=1}^{J} \eta_{jj},$$

where η_{jj} represents the proportion of examinees falling into the same j^{th} observed and true performance category.

Method

Data

To make simulated data as realistic as possible, item parameters used for generating the simulated data were obtained by calibrating AP datasets. Specifically, nine AP examinations administered in 2012 were used for calibration including Spanish Language and Culture, Spanish Literature, French Language and Culture, Italian Language and Culture, English Language and Culture, German Language and Culture, World History, Biology, and Environment Science. All the examinations are mixed-format tests consisting of multiple-choice (MC) and free-response (FR) items. Thus, two distinct item pools were created for MC and FR items, respectively. A summary of item pool information is provided in Table 1. The three-parameter logistic model (Lord, 1980) and the graded response model (Samejima, 1997) were used for fitting the dichotomously-scored and polytomously-scored data, respectively, using *flexMIRT* (Cai, 2012). The estimated item parameters for each types of item were put onto either the MC or FR item pool.

Study Design

In this section, the designs of three studies are illustrated and the data generation procedure is discussed. Each simulation study was designed to produce a unique observed score distribution. Test length was fixed to forty for all three studies (i.e., the number of possible score points is forty one, including zero score). It is important to note that examinees were assumed to be fixed and items to be random in this study. In other words, for each replication the same set of examinee ability parameters drawn from a particular distribution were used for generating data based on a different set of item parameters. The ability parameters were also used to define the criterion classification indices as discussed in a later section.

Study 1. The first study investigated the performance of estimation procedures under a bimodal distribution. In Study 1, all forty items were MC items, with a score range of 0 - 40. Bimodal distributions for θ were generated by combining two normal distributions with different mean values of θ (i.e., Normal (-1.8, $\sqrt{.8}$) and Normal(.8, $\sqrt{.8}$)). Figure 1 depicts the shape of generated observed scores, and is discussed further at the end of this section.

Study 2. The second study explored the effects of structural bumpiness in a score distribution. A test consisted of four FR items and each item was scored 0-10 with a non-consecutive increment (i.e., 0, 1, 4, 5, 8, and 10), resulting in a score range of 0 - 40. Due to the non-consecutive increment in an item score, the generated score distribution was expected to have structural bumpiness. Note that structural bumpiness indicates some systematic bumpiness in a score distribution resulting from some types of scoring rules or test specifications such as section weights and is not caused by sampling error.

Study 3. The last study examined the impact of structural zeros in a score distribution, which are score points that cannot occur and thus have zero frequencies. For Study 3, a mixed-format test was generated. The weights for the MC and FR sections were intentionally chosen to create observed score distributions having structural zeros (i.e., using weights of 2 for each section). Unlike the previous two studies, using double weights for each section led to a score range of 0 - 80.

Once the data generation process was complete, it was necessary to inspect consistency between the targeted distributions and the distributions actually generated to ensure that they were created as intended. However, the simulation procedure used in this study posed a challenge in defining the population distributions, because a different set of items was drawn

from the item pools for each replication. Having different items in a test form necessarily leads to different shapes of a score distribution for each replication. Thus, instead of looking at score distributions from every replication, a score distribution was obtained using a large sample size of 100,000, which is displayed in Figure 1. In Figure 1, a blue line represents the bimodal score distribution designed for Study 1; an orange line shows the distribution with structural bumpiness for Study 2; and a gray line indicates the distribution with structural bumpiness generated for Study 3. In this figure, it is evident that the generated distributions for all three studies reflect their intended unique characteristics. It is important to note that Figure 1 was not actually used in the analyses. Rather, it serves as a tool for verifying the intended score distributions.

Simulation Conditions

Cut scores were arbitrarily chosen at the 50%, 65%, and 80% of the maximum possible score. Results were obtained for binary classifications by applying these cut scores one at a time as well as for multi-level classifications by applying all three cut scores simultaneously. Three levels of sample size were considered for each study: 100, 1,000, and 5,000.

Analysis

NM procedure. A requirement for conducting the NM procedure is to compute a reliability coefficient. In this study, the reliability coefficient was estimated using the CM model (Lee, 2007), which roughly can be viewed as a KR-21 analogue for non-dichotomous items. The computer program R was used to implement the NM procedure.

LL procedure. Implementation of the LL procedure requires four pieces of information: (a) an observed score distribution, (b) a reliability coefficient of the scores, (c) maximum and minimum possible scores on the test, and (d) cut scores. Similar to the NM procedure, the reliability coefficient was obtained using the CM model. For the analysis, the computer program BB-CLASS (Brennan, 2004) was used.

CM procedure. The program MULT-CLASS (Lee, 2005b) was used to execute the CM procedure. Brennan and Lee (2006) proposed a bias-correction procedure as an effort to reduce bias resulting from the use of the observed proportion correct score as an estimate of each person's true proportion correct score. An extension of the bias-correction procedure for mixed-format tests was made by Wan et al. (2007). Both studies reported a substantial reduction in bias

in classification consistency estimates by using the proposed procedures (Brennan & Lee, 2006; Wan et al., 2007). Thus, this study employed the bias-correction procedure for the CM method.

Evaluation Criteria

Root mean square error (RMSE), absolute value of bias (ABS), and standard error (SE) were computed to evaluate each procedure. Denoting α as a parameter of interest (i.e., either classification consistency or accuracy index), the ABS, SE, and RMSE of an estimator of α are expressed as:

$$ABS(\hat{\alpha}) = \left|\bar{\hat{\alpha}} - \alpha\right|,$$

$$SE(\hat{\alpha}) = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\alpha}_i - \bar{\hat{\alpha}})^2},$$

and

$$RMSE(\hat{\alpha}) = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\alpha}_i - \alpha)^2},$$

where r is the number of replications, which was set to 100 in this study.

The criterion values of classification indices (i.e., α) were obtained for each pre-specified sample size based on a sample of simulated examinees. Examinees were assumed to take a pair of parallel forms. The parallel forms of a test were constructed by randomly drawing items from the MC and FR item pools. This procedure was repeated 1,000 times. The specific steps for determining criterion classification consistency indices are as follows:

1. create two item pools for MC and FR sections, respectively;
2. construct two test forms by randomly drawing items from the item pools;
3. generate item responses of each examinee for the two forms using the known item and ability parameters. Here, ability parameters are identical to those used for generating data for each replication;
4. for each form, classify the examinees into categories based on the observed scores;
5. compute classification consistency indices; and
6. repeat Steps (3) to (6) 1,000 times and compute the average of classification consistency indices, which were regarded as the criterion values.

For the criterion classification accuracy, the expected score over two forms and over 1,000 replications (i.e., 2,000 parallel forms) was regarded as the true score for an examinee. In Step (3), examinees were categorized based on their true score and the observed score (note that only one observed score on one form was used and the other form was disregarded). The average value of classification accuracy estimates over 1,000 replications was considered as the criterion classification accuracy.

The criterion indices were obtained for each sample size condition for each study. Note that the criterion indices were obtained under the assumption of randomly parallel forms, which is generally consistent with the underlying assumption of the LL and CM methods, but not necessarily with the NM method. Therefore, it is possible that the criteria might favor the LL and CM methods over the NM method to some extent.

Results

In this section, the results are presented for each of the three studies, and each study is discussed in terms of *P*, kappa, and gamma indices. Note that SE reflects the amount of random error in the estimates, whereas ABS represents the magnitude of the systematic error introduced by an estimation procedure. RMSE quantifies the overall error involved in the estimates.

Study results can be seen in Figures 2 through 10. The first three figures pertain to the results for Study 1, the next three figures to Study 2, and the last three to Study 3. Among three figures for each study, the first figure includes the results for the *P* index, the second for the kappa coefficient, and the third for the gamma index.

For each figure, each column represents a sample size condition. The first row contains ABS results for the three estimation procedures, the second row contains SE results, and the last row includes RMSE results. In each plot, the horizontal axis represents three binary cut score points: cut 1, cut 2, and cut 3 for 50%, 65%, and 80% of the maximum possible score, respectively, along with the simultaneous (multi-level) classification.

Study 1: Bimodal Distribution

Agreement index *P*. Regarding the agreement index *P*, the results for Study 1 with a bimodal distribution are provided in Figure 2. In general, the NM procedure tends to have larger bias than the other two procedures, which consequently leads to larger RMSE values for the NM. The NM procedure introduces a considerable amount of bias for multi-level classifications across all sample size conditions. The magnitude of bias tends to decrease at cut 3, at or below which

approximately 75 percent of examinees are located. This result suggests that there is a higher degree of consistency between the model-based normal and actual distributions at cut 3. The increase in the sample size does not seem to reduce the magnitudes of bias, but rather it reduces SE values for all estimation procedures.

The differences in SE values among estimation procedures become smaller as the number of examinees increases as can be seen in the second row of Figure 2. As a result, all three procedures provide very similar SE values under large sample size conditions. Overall, the smallest SEs tend to be associated with the NM procedure, although the differences among procedures do not exceed more than .015 in any conditions.

The smallest RMSE tends to be found for the CM procedure with the small sample size condition (i.e., N=100). For the larger sample size conditions (N=1,000 and 5,000), the CM procedure performs best in the cut score locations of 65% and 80%, but the LL procedure produces slightly smaller bias and RMSEs than the CM on the 50% cut score and simultaneous classifications. It is surprising that the LL procedure shows satisfactory performance under the bimodal score distribution, alleviating the potential concerns about its potential inaccuracy with a bimodal shape of scores. It is most evident that the NM method produces substantially larger error than the other procedures under the bimodal distribution, primarily due to its large bias.

Kappa coefficient. The results for study 1 for kappa coefficient are plotted in Figure 3. In general, all three procedures have larger bias and SE values for kappa than for P, which has frequently been reported in the literature (Wan et al., 2007; Wan, 2006). However, the general pattern of each estimation procedure is very similar to the results for P, with a small exception of the NM results. The relative performance of the NM procedure is slightly improved for cut 3. This is mainly due to its substantially small value of SE at cut 3.

Gamma index. The results for the gamma index for study 1 can be found in Figure 4. The patterns for the three estimation procedures are very similar to those for the P results. However, the magnitudes of overall error in the gamma index are generally smaller than those in P and kappa. With regard to bias, the NM procedure tends to yield the largest error, followed by the LL procedure, and then by the CM procedure. However, in terms of SE, NM always produces the smallest SE values, followed by CM, and next by LL. Overall, the smallest RMSE tends to correspond to the CM procedure primarily due to small bias.

Study 2: Distribution with Structural Bumpiness

Agreement index *P*. In study 2, the performance of three estimation procedures was investigated under the distributions having structural bumpiness. The results for agreement index *P* can be seen in Figure 5. In terms of bias, the largest is generally associated with the CM procedure with an exception at cut 3, where CM shows the least bias except for the sample size of 5,000. For the multi-level classifications, LL is associated with the lowest bias, while CM has the largest bias. One noteworthy finding is that the LL procedure tends to be affected substantially by the increase in sample size—bias approaches zero when sample size is 5,000. LL, among the three methods, is the only method that involves estimation of the true score distribution for the population of examinees, and a larger sample size would lead to more accurate estimation. Note also that the effective test length is used for LL to define a conditional distribution for errors, which, consequently, has an impact on the marginal distribution and estimated classification. The effective test length is determined as a function of sample statistics such as the minimum, maximum, mean, etc. Increasing sample size likely improves the accuracy of these statistics. A large sample size is generally preferred for all estimation procedures; however, the LL procedure appears to be much more sensitive to the sample size. This pattern is observed across all three studies. Further investigation needs to be conducted regarding the relationship between sample size, effective test length, and the accuracy of classification indices.

With respect to SE, the LL procedure provides the largest values, followed closely by CM, and NM with the exception of multi-level classifications. For multi-level classifications, CM has the largest SE although the difference with the other procedures is very small. As discussed in study 1, the NM procedure almost always results in the smallest SE. However, the difference among three procedures decreases as the number of examinees increases. The increase in sample size tends to reduce the SE, which is reasonable given that random error is inversely proportional to sample size.

The pattern of RMSE closely follows the pattern of bias. Overall, the CM performs poorly compared to the other procedures, and the NM procedure generally provides the reasonably accurate estimates. The performance of LL improves with a large sample size.

Kappa coefficient. The results for kappa coefficient can be seen in Figure 6. In general, the NM procedure has smaller bias than the CM procedure across all sample sizes. For those two procedures, the magnitudes of bias seem to be consistent over all cut score locations. The LL

procedure shows a considerable amount of bias at cut 3 compared to the other cut score positions. Also, as in the case of *P*, the LL procedure tends to have lower bias as the sample size increases.

The largest SEs are mainly associated with the LL procedure across all sample sizes, which is a consistent finding with the results for *P*. The lowest SEs always correspond to the NM procedure. One of the interesting findings is that the CM procedure leads to larger SE than the LL procedure only for multi-level classifications. This tendency was also found in the *P* results.

Concerning the RMSE, the NM procedure always shows the smallest RMSE values across all study conditions including sample size and position of cut scores. When sample size is large, the CM procedure generally shows higher error than the other two procedures, with the exception of cut 3 where the LL method has the largest error. Under a small sample size condition (*N*=100), the LL procedure has the largest amount of RMSE, except for multi-level classifications. The similar tendency was also observed for the *P* results.

Gamma index. The results for the gamma index are summarized in Figure 7. The general trend of each procedure is similar to that for the *P* results, while the behavior of the LL procedure somewhat differs. Specifically, regardless of sample size, the LL procedure, compared to the NM procedure, tends to produce larger error, except for cut 1. Moreover, the LL procedure consistently provides larger bias than NM at the 80% cut score. Similar to the *P* results, Figure 7 suggests that the increase in sample size helps reduce bias substantially for the LL procedure.

Often, the *P* index has lower values than the accuracy index because the former is determined by two observed scores introducing measurement error twice, whereas the latter involves measurement error from a single observed score only. This relationship has been reported by previous studies (Haertel, 2006; LaFond, 2014). However, as can be seen in Figure 7, there are some cases where the errors in the accuracy estimates are slightly larger than the errors in the *P* estimates mainly due to the larger SE values, especially for the LL procedure with a small sample size condition (*N*=100). As with SE, the smallest SEs correspond to the NM procedure, followed next by the CM, and last by the LL procedures.

In terms of RMSE, when sample size is small, the NM procedure has the lowest error, followed by CM. The difference between the LL and CM procedures becomes smaller with the increase in sample size. Note that as the number of examinee increases, the accuracy of the NM procedure on the 50% cut score point becomes relatively worse.

Study3: Distribution with Structural Zeros

Agreement index *P*. Study 3 was designed to evaluate the accuracy of the selected procedures when a score distribution involves structural zeros. The results for the agreement index *P* are depicted in Figure 8. In terms of bias, the pattern of the three estimation procedures varies depending on sample size. That is, the relative performance between the procedures reveals a mixed picture across different sample size and cut score positions. Specifically, it appears that the LL procedure consistently produces small bias, keeping the magnitudes of bias near .01 over sample size and cut score location conditions. The NM procedure tends to produce larger bias with a small sample size (*N*=100), while the CM procedure tends to show larger bias under the large sample size conditions, in general. However, their relative performance is highly dependent on the position of the cut score. Therefore, the results do not offer a clear pattern across various sample size and cut score location conditions.

As with Study 1 and Study 2, the NM procedure tends to provide the smallest SE, followed by the CM, and by the LL procedures. This general pattern does not change much over the sample size conditions.

As with the case for bias, no clear tendency could be found with the RMSE results because the RMSE patterns closely follow the bias patterns. The accuracy of each procedure irregularly changes with different conditions of sample size and cut score standing. When simultaneous multi-level classifications are of primary interest, however, the smallest RMSEs are always associated with the NM procedure across all sample size conditions.

Kappa coefficient. The results for the kappa coefficient can be seen in Figure 9. A quick inspection of Figure 9 reveals that kappa behaves very differently from *P*. Also, the accuracy results show a somewhat unique pattern, as will be discussed later.

In Figure 9, in general, the RMSE values of the NM procedure are fairly comparable to the CM procedure. Both procedures tend to consistently stay within a range of .04-.07. On the other hand, the LL procedure seems to introduce a large amount of error particularly for cut 3 when sample size is small (*N*=100). The accuracy of the LL procedure improves with increasing sample size.

Gamma index. The performance of each estimation procedure concerning the gamma index can be found in Figure 10. As noted earlier, the results quite differ from the *P* or kappa results, which is not consistent with the conventional belief. Another notable pattern is that the

LL procedure introduces a considerable amount of bias across most of the conditions. This tendency becomes less visible as sample size increases although its bias is still larger than the others. The CM procedure tends to work reasonably well particularly for the small sample size condition. As sample size increases, the magnitudes of bias becomes comparable between the CM and NM procedures. In sum, the LL procedure consistently reveals substantial errors regardless of sample size compared to the other procedures, and CM slightly outperforms NM with a small sample size but their difference becomes less distinguishable as sample size increases.

Discussion

When a test is designed to classify examinees into multiple categories (e.g., pass or fail), the classification decision usually has high-stake consequences such as graduate/license requirements or school accountability. For such tests, classification consistency and accuracy indices are important psychometric properties for interpretations made based on these test scores. Although a number of procedures for estimating classification indices have been evaluated in terms of their relative performance, little research exists in the literature that examines their performance in relation to various types of score distributions. This study is aimed to fill the gap in the literature.

Three non-IRT estimation procedures were investigated including the normal approximation (Peng & Subkoviak, 1980), Livingston and Lewis (1995), and compound multinomial (Lee, 2005) procedures. The selected procedures are based on different score distribution assumptions employing different psychometric models. Therefore, it was anticipated that they would show different patterns under various distributions. The distributions considered in this study involve a bimodal distribution, a distribution with structural bumpiness, and a distribution having structural bumpiness. Selection of distributions was intended to reflect the potentially plausible score distributions, although they may not be seen very often in practice. A series of simulations was carried out to investigate the performance of the estimation procedures under the three atypical distributions.

The results of this study led to the following conclusions. First, under a bimodal distribution, NM introduces substantial bias, whereas both CM and LL provide smaller error. This conclusion is true for all three classification indices. Second, for a distribution with structural bumpiness, CM tends to have significantly larger bias, and in general, NM provides

lower error than LL although the difference between NM and LL becomes less noticeable as sample size increases. Third, the largest SE consistently corresponds to LL across all study conditions, with a few exceptions. Also, NM tends to be associated with the smallest SE. Similar patterns were also observed in Wan et al. (2007). Fourth, under a distribution with structural zeroes, no definite statements in regard to the performance of the procedures could be made because their relative performance varies across the cut score location and sample size conditions. Also, the results differ depending on which classification index is used. For the *P* index with simultaneous multi-level calibrations, CM tends to produce larger error than NM and LL across all sample size conditions. By contrast, for the kappa coefficient and gamma index, LL results in larger error than the other two across most of the conditions. Fifth, LL appears to be more sensitive to sample size than the other two methods. As sample size increases, reduction in error is most noticeable with LL. Last, all three estimation procedures perform reasonably well in recovering true classification indices when applied to score distributions that are not consistent with the model-based distributions. The amount of bias for all indices observed in this study, on average, was about .02~04. However, there were cases where bias exceeded .10.

Overall, it seems prudent to conclude that the user consider factors such as the type of score distribution, sample size, and cut score location when selecting an estimation method in practice. It is also important to consider the unique assumptions about test form parallelism and data structure that each estimation procedure possesses. Another caution is that the significance of the findings of this study should not be exaggerated. It is worth mentioning that the maximum differences among estimation procedures were around .14 in magnitude and could be considered to be small in practice. The significance of these differences highly depends on the purpose of a test and test users' interpretations.

Limitations and Future Research

Several limitations of the study should be acknowledged. This study was conducted based on simulations, which might not exactly reflect data structure observed in real data. Caution should be made in generalizing the results. Therefore, a future study should involve real data analysis. Second, only non-IRT procedures were investigated in this study. The performance of other estimation procedures such as IRT procedures might also be examined in the future study. Last, the criterion classification indices were obtained for each sample size separately based on the assumption that the criterion is an average value of classification indices for a

specific group of examinees taking an infinite number of randomly parallel forms of a test. However, the criterion values can be established in other ways with different assumptions. For example, a criterion can be regarded as an average value of classification indices for a very large number of examinees taking an infinite number of randomly parallel forms. With this criterion, only a single criterion value could be used across different sample size conditions. Different criteria, of course, would lead to different study results. A future study might include various criteria under different assumptions and explore the differences among the study results resulting from the use of a different set of criteria.

References

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy, Version 1.1* (CASMA Research Report No. 9). Iowa City, IA: University of Iowa.

Brennan, R. L., & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City, IA: University of Iowa.

Cai, L. (2012). *flexMIRT* (Version 1.88) [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Deng, N. (2011). *Evaluating IRT- and CTT- based method of estimating classification consistency and accuracy indices from single administrations.* Unpublished Doctoral Dissertation, The University of Massachusetts, Amherst.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Westport, CT: Praeger Publishers.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253-264.

LaFond, J. L. (2014). *Decision consistency and accuracy indices for the bifactor and testlet response theory models*. Unpublished Doctoral Dissertation, The University of Iowa.

Lee, W. (2005a). *Classification consistency and accuracy under the compound multinomial model*. (CASMA Research Report No. 13). Iowa City, IA: University of Iowa.

Lee, W. (2005b). *Manual for MULT-CLASS: For multinomial and compound multinomial classification consistency*. Iowa City, IA: University of Iowa.

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31,* 255-274.

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

Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33*, 374-390.

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

Li, S. (2006). *Evaluating the consistency and accuracy of proficiency classifications using item response theory.* Unpublished doctoral dissertation, The University of Massachusetts, Amherst, MA.

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: Erlbaum.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

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

The College Board. (2016, December 21). *AP Students – AP Courses and Exams for Students.* Retrieved from https://apstudent.collegeboard.org/home

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). Iowa City, IA: University of Iowa.

Wan, L. (2006). *Estimating classification consistency for single-administration complex assessments using non-IRT procedures.* Unpublished Doctoral Dissertation, The University of Iowa.

Table 1

Item Pool Information

	Score Range	# of Items	Item Parameters	Range of Item Parameters	Mean of Item Parameters	SD of Item Parameters
MC Item Pool	0-1	657	*a*	.108 ~ 2.235	.8155	.3162
			b	-4.745 ~ 3.0217	-.3116	1.0484
			c	.0226 ~ .5452	.1833	.0999
FR Item Pool	0-10	8	*a*	.9577 ~ 1.2195	1.0805	.1042
			*b*1	-2.3404 ~ -1.5463	-1.8350	.3143
			*b*2	-1.5311 ~ -.7330	-1.1473	.3093
			*b*3	-.8770 ~ -.1012	-.5477	.3161
			*b*4	-.4364 ~ .4628	-.0192	.3287
			*b*5	-.0242 ~ .9923	.4537	.3647
			*b*6	.4140 ~ 1.5239	.9242	.4098
			*b*7	.8410 ~ 2.0878	1.4036	.4808
			*b*8	1.3209 ~ 2.7192	1.9261	.5733
			*b*9	1.8451 ~ 3.5369	2.5538	.7155
			*b*10	2.5148 ~ 4.2761	3.2736	.7826
	0-5	21	*a*	.6361 ~ 1.7053	1.0168	.3232
			*b*1	-5.5202 ~ -1.9127	-2.9879	.8873
			*b*2	-3.5113 ~ -1.0895	-2.0399	.7638
			*b*3	-2.8014 ~ .0576	-.9828	.7588
			*b*4	-1.2324 ~ 1.5885	.2848	.7437
			*b*5	.2988 ~ 3.0753	1.4529	.7953

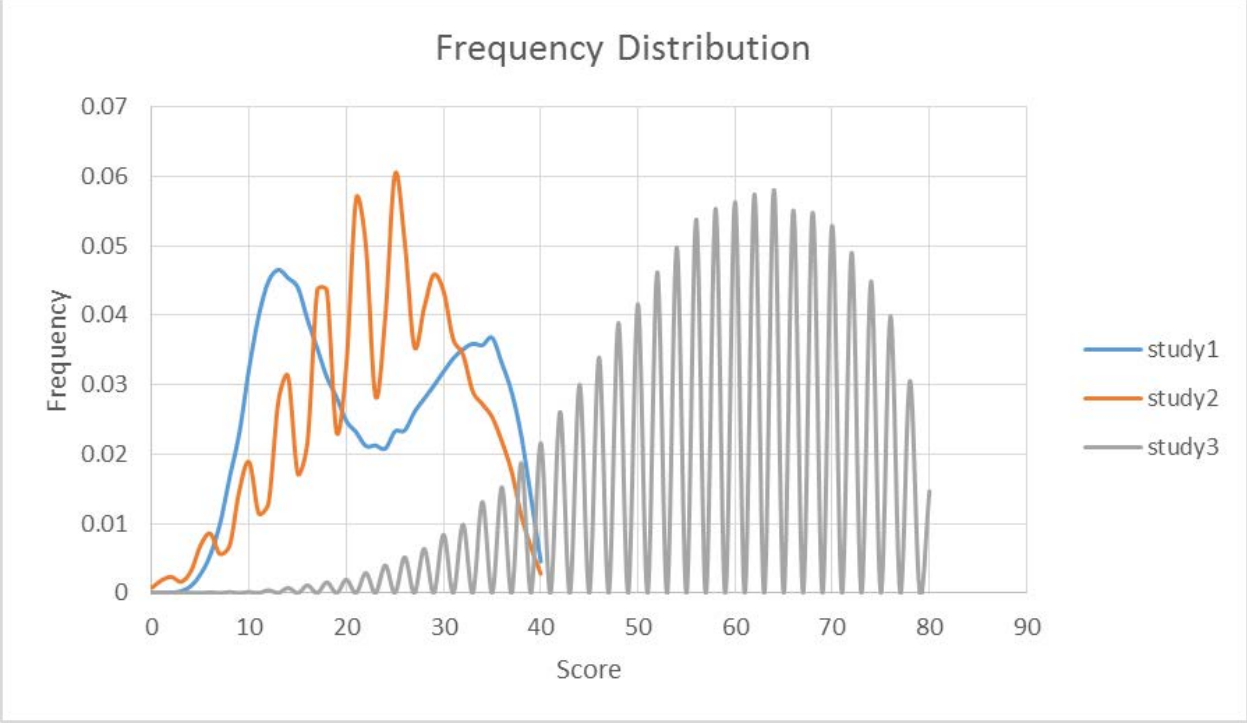

Figure 1. Frequency distributions based on a large sample size for each study.

Figure 2. Results of agreement index *P* for Study 1 (bimodal distribution).

Figure 3. Results of kappa coefficient for Study 1 (bimodal distribution).

Figure 4. Results of accuracy index (γ) for Study 1 (bimodal distribution).

Figure 5. Results of agreement index *P* for Study 2 (distribution with structural bumpiness).

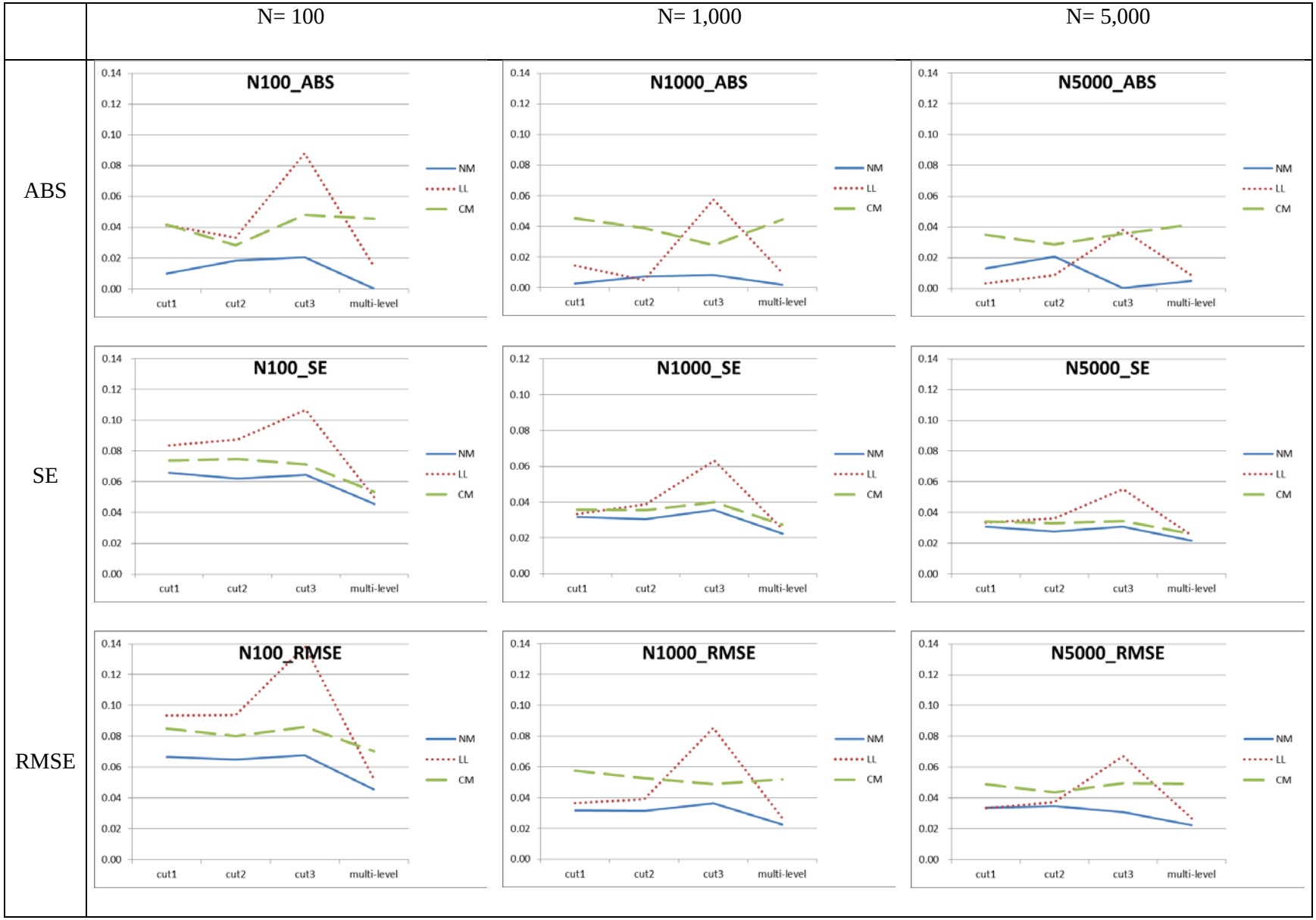

Figure 6. Results of kappa coefficient for Study 2 (distribution with structural bumpiness).

Figure 7. Results of accuracy index (γ) for Study 2 (distribution with structural bumpiness).

Figure 8. Results of agreement index *P* for Study 3 (distribution with structural zeros).

Figure 9. Results of kappa coefficient for Study 3 (distribution with structural zeros).

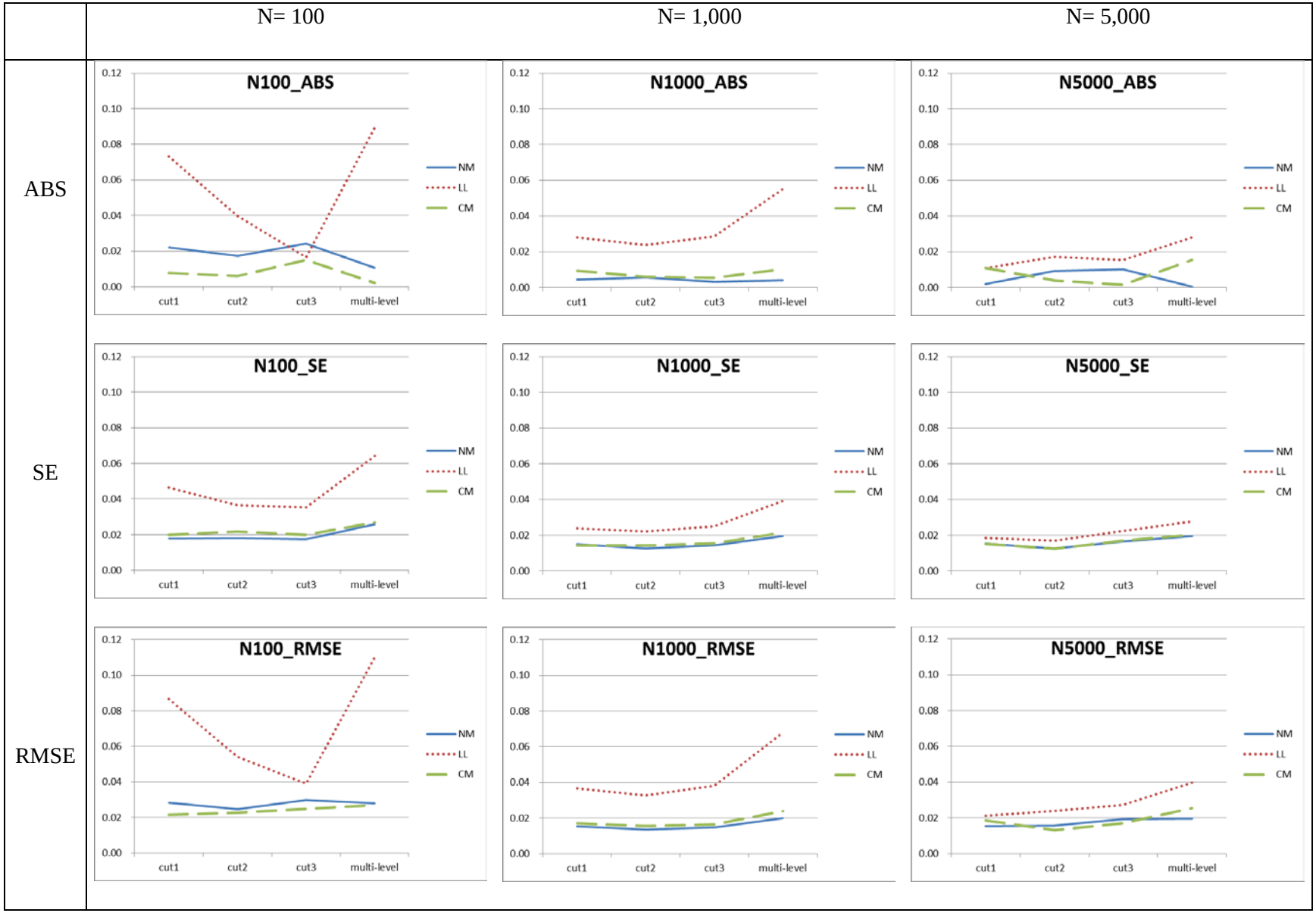

Figure 10. Results of accuracy index (γ) for Study 3 (distribution with structural zeros).

Chapter 7: Reliability of Mixed-Format Composite Scores Involving Raters: A Multivariate Generalizability Theory Approach

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

Abstract

This study examines various multivariate generalizability theory designs that can be used for mixed-format tests. In particular, this study focuses on composite score reliability estimates considering potential sources of error associated with each item format including a rater facet. Several D studies were carried out with various numbers of items per item-format section and two levels of the number of raters. The study results showed that the error variance and reliability estimates can be biased if there is a mismatch between the universe of admissible observations and the structure of available data**.**

Reliability of Mixed-Format Composite Scores Involving Raters: A Multivariate Generalizability Theory Approach

Reliability is one of the most important psychometric properties, which is often evaluated to quantify the precision of measurement. A great deal of research has been devoted to developing frameworks for estimating reliability. Generalizability theory (G theory) can be regarded as one of the fundamental psychometric frameworks developed to accomplish this goal (see Brennan, 2001a). However, very little research exists for estimating reliability for mixed-format tests using G theory, or any other theory for that matter.

Mixed-format tests are usually defined as having two types of item format: multiple-choice (MC) and free-response (FR). Estimating reliability for mixed-format tests can be complicated compared to estimating reliability for single-format tests because having multiple item formats introduces multiple sources of error. For example, FR items are often scored by human raters, which may potentially introduce errors attributable to rater variability. The rater effect, however, usually is not considered for MC items. Despite its complexity, the increasing popularity of mixed-format tests calls for more accurate reliability estimates by taking into consideration potential sources of error related to mixed-format tests. This study uses real data to illustrate how to obtain more defensible estimates of reliability for mixed format tests.

A very limited number of studies have been conducted to estimate reliability for mixed-format tests using G theory. Much of the previous research that used G theory has focused primarily on performance assessments or portfolio assessments consisting of a single type of item format (Gao & Brennan, 2001; Jarjoura, Early, & Androulakakis, 2004; Clauser, Balog, Harik, Mee, & Kahraman, 2009). One of the first attempts to apply G theory to mixed-format tests was made by Powers and Brennan (2009). They examined the effects of various factors associated with mixed-format tests (e.g., section weights and the numbers of items per section) on composite score reliability, error variance, and conditional errors of measurement. The G theory design that they used was a multivariate design with items nested within an item format facet. In a recent study, Moses and Kim (2015) proposed an extended version of the existing multivariate design that can be used for mixed-format tests. They also provided a simulation procedure for estimating classification consistency and classification accuracy indices within the G theory framework.

For those two previous studies, some limitations were acknowledged by the authors. Powers and Brennan (2009) did not include a rater facet because only a single rating of the response of each person was available for FR items. As a result, they admitted that FR score reliability based on their study design may be at least somewhat inaccurate. A similar limitation applies to Moses and Kim (2015). In this study, although double scoring was achieved for each FR item, raters were not systematically assigned. Consequently, the data collection design precluded more desirable G theory designs such as a crossed or nested rater effect designs. Recognizing the limitations of previous research, the present study is intended to estimate composite score reliability using multivariate G theory while considering the potential sources of error associated with each item format including a rater facet. Several D studies were carried out with various numbers of items per section and two levels of the number of raters.

Method

General Description of Exam

The AP German Language exam administered in 2013 was considered in this study. This exam is composed of MC and FR sections. Operationally, composite scores are defined as a weighted sum of the MC and FR section scores with a set of pre-specified section weights. The section weights are established a priori by test developers in order to achieve the intended proportion of points for each section based on test specifications. We will use the term "relative weights" to refer to the proportional contributions of the section scores to the composite score, which is consistent with the terminology used by Powers and Brennan (2009). For the data used in this study, the relative weights are equal to .5 for both MC and FR sections.

The exam has 65 MC items scored 0 or 1, and 4 FR items scored 0 - 5. Consequently, the nominal weight (see Brennan, 2016a), which serves as a multiplier when computing composite scores, is 1 for MC items and 3.25 for FR items (i.e., score of 1 for a FR item is equivalent to 3.25 MC points). This weighting scheme leads to each section score contributing 50% to the composite score that ranges from 0 to 130 (2×65). These nominal weights are defined for the total score metric, which is seldom used in G theory. Within the G theory framework, the usual convention is to use the mean score metric. Based on the mean score metric, the composite score (C) can be defined as

$$w_M \bar{X}_M + w_F \bar{X}_F = C,$$

where w_M and w_F indicate the nominal weights for the MC and FR sections, respectively, and $\bar{X}_M$ and $\bar{X}_F$ are the mean scores for the MC and FR sections. Based on this formula, the nominal weights for MC and FR items on the mean score metric were 65 and 13 (4×3.25), respectively. For more information on how to determine various sorts of weights, readers might refer to Brennan (2016a).

Study Designs

For the multivariate designs used in this study, the structure of the universe of admissible observations (UAO) conceptually follows a multivariate $p^\bullet \times i^\circ \times r^\bullet$ design with item format serving as a fixed multivariate variable consisting of two levels, MC and FR item types. A unique feature of this design is that raters are only associated with FR items, not MC items. That is why a half-closed circle is used to designate the rater facet. For the MC section, alone, the structure of the UAO is $p \times i$; for the FR section, alone, the structure is $p \times i \times r$. In this multivariate UAO, persons (p) are crossed items (i) since any item can be taken by any person in the population. The closed circle associated with persons indicates that any person in the population takes both item types, MC and FR. Items (i) are always nested within item types denoted by an open circle. For the FR items, it would be reasonable to regard items (i) being crossed with raters (r) in the UAO, as opposed to being nested, because any particular FR item could be scored by any particular rater.

Operationally, each FR item is scored by a single rater, which makes it impossible to disentangle variability attributable to the two different sources, raters and items, for FR scores. Note also that even though the UAO has a crossed design for the rater and item facets, the G study design based on the operational data may fail to reflect this structure because different raters are assigned to different FR items (i.e., raters are nested within items). Thus, three different multivariate G study designs were considered in this study: $p^\bullet \times i^\circ$, $p^\bullet \times i^\circ \times r^\bullet$, and $p^\bullet \times (r^\bullet : i^\circ)$. A description of each design is provided next. Note that for all three multivariate designs, item format is a fixed facet, and MC and FR are the two conditions of this facet.

$p^\bullet \times i^\circ$ design. The $p^\bullet \times i^\circ$ design is often used when there is a distinct set of items nested within each of the levels of a fixed facet, such as content categories (Brennan, 2001a). Since each person took all items in both the MC and FR sections, persons are fully crossed with the item format facet (designated with a closed circle or bullet), while items are nested within the item format facet (designated with an open circle).

This design does not involve raters as a facet, and is typically used for analyzing the operational data where only a single rating is available for each FR item. Since the detailed information about the actual rating procedure was not available for the operational data used in this study, it was "assumed" that a single rater evaluated all FR items for all examinees. If the same rater was used for all FR items and all examinees, this design effectively treats raters as a *hidden* and *fixed* facet. Recall that the UAO has a fully crossed design involving both *i* and *r*; however, in the $p^{\bullet} \times i^{\circ}$ design, raters are hidden and cannot be separated from the other facets. When there is a mismatch in designs between the UAO and the available data and/or only a single condition of a facet is sampled, several problems arise such as interpretational complexities and bias in some statistics (see Brennan, 2016b). Specifically, for the FR section, the rater facet is confounded with the item and person facets, which can be expressed in terms of the random effects variance components for the fully crossed design as:

$$\sigma^2(p|r) = \sigma^2(p) + \sigma^2(pr),$$
$$\sigma^2(i|r) = \sigma^2(i) + \sigma^2(ir),$$
$$\sigma^2(pi|r) = \sigma^2(pi) + \sigma^2(pir),$$

where "$|r$" indicates that the rater facet is a single fixed condition. Confounding will affect the error variances and reliability coefficients in subsequent D studies.

However, alternatively, the rater facet could be considered as a hidden *random* facet, if different raters were used to score different FR items. In this case, raters are completely confounded with items (see Brennan, 2016b), and the D study results will be affected by the confounding. It is important to remember is that the rater facet is not specified explicitly in the $p^{\bullet} \times i^{\circ}$ design, which would produce D study statistics such as error variance and generalizability coefficient that differ from the ones based on the fully crossed design. Note that the fully crossed design can be considered "ideal" in the sense that it mirrors the structure of the UAO.

$p^{\bullet} \times i^{\circ} \times r^{\circ}$ design. The second design is rarely seen in the literature. This multivariate design is somewhat similar to Moses and Kim's approach (2015). They provided a multivariate G theory design that can be used for mixed-format tests, which was denoted as $p^{\bullet} \times (i^{\circ}\ or\ j^{\circ}) \times h^{\circ}$ in their paper. With their notation, the *i* and *j* items (*i* for FR, *j* for MC) are nested within item format, whereas the rating (*h*) facet is defined only for FR items. In their study, ratings, not raters, were used as a facet due to the limited data collection design. By contrast, in the present

study, data from a special study allowed us to further investigate the rater effects, keeping the general study design virtually identical to the design used in Moses and Kim (2015).

This second multivariate design is denoted here as $p^{\bullet} \times i^{\circ} \times r^{\diamond}$. The half-closed circle emphasizes that the rater facet is partially crossed with the fixed item-format facet in the sense that the rater facet occurs for the FR section, only. Note that the raters are crossed with the FR items in this design.

In this multivariate design, the $p \times i$ design was employed for the MC section, while $p \times i \times r$ design was used for the FR section. The reason that two separate analyses were done is because the two sections differ in the structure of the UAO. In other words, because the FR section is composed of several items each scored by different raters, variability in both raters and items contributes to error variance in test scores. On the other hand, for the MC section, variability associated with raters does not occur.

$p^{\bullet} \times (r^{\diamond}: i^{\circ})$ design. The last study design is similar to the second design in that the rater facet involves a half-closed circle; however, this last design is different from the second design since the raters are nested within items as opposed to being crossed. Again, the half-closed circle associated with the rater facet implies that it only exists for the FR section.

For the nested design, some variance components are confounded when compared to those for the fully crossed design. Specifically,

$$\sigma^2(r{:}\,i) = \sigma^2(r) + \sigma^2(ir),$$
$$\sigma^2(pr{:}\,i) = \sigma^2(pr) + \sigma^2(pir),$$

which leads to smaller error variances and larger reliability coefficients than the crossed design. If the UAO has a crossed design, reliability will be overestimated with the nested design, all other conditions being equal (Brennan, 2001a).

Data Structure

Three sets of data were obtained: one from the operational administration and the other two from a special study conducted by the College Board. Because, operationally, an examinee's response to a FR item is evaluated by a single rater, the effect of rater variability typically cannot be studied using operational data. A special study was conducted to obtain ratings for each FR item from multiple raters so that the rater effect could be incorporated in a reliability analysis.

For the special reader reliability study, FR answer books were selected from the operational administration (N=112) and scored by two independent raters. Data from the

operational administration and the special study were matched by the examinees' registration identification numbers to identify the 112 examinees. Examinees with missing responses to any FR item or more than 5 MC items were dropped, resulting in the total sample size of N=109. Not reached or omitted responses to the MC items were scored as incorrect. As a result, each of the 109 examinees had three scores for each FR item—one operational score plus two additional scores from the special study. Figure 1 depicts the structure of the three datasets. As shown in Figure 1, there were three scores for each FR item for each examinee, as marked with ②,③, and ④. For the special study, two separate datasets were defined such that each dataset was associated with FR scores rated by one of the two raters (③ or ④).

For the special study, not all examinees were rated by the same raters so that the facets (i.e., examinees, items, and raters) were not fully crossed. The data structure for the special study is given in Figure 2. Again, note that the data for the special study were collected only for the FR items as presented in Figure 1. As explained previously, each FR item (i) for each examinee (p) was scored by two independent raters (r), and there were a total of 4 FR items, 109 examinees, and 8 raters. In Figure 2, the same shaded pattern indicates the same pair of raters. Eight folders were distinguished based on the combination of item, examinee, and rater pair. Note that for each folder, a pair of raters scored a FR item, so each of the two scores is associated with either dataset ③ or ④ in Figure 1. Three G theory designs were considered in this study, each of which involved a different combination of data as explained below.

$p^{\bullet} \times i^{\circ}$ design. Since there were three FR scores available for each FR item based on the operational administration and special study (i.e., ②,③, and ④ in Figure 1), three separate G-study analyses were conducted to obtain three sets of estimated G-study variance and covariance components(i.e., ①+②, ①+③, and ①+④ in Figure 1). The averaged values over the three analyses were treated as final estimates, which were then used as an input to obtain D study results. mGENOVA (Brennan, 2001b) was used to estimate the G study variance and covariance components, D study variance and covariance components, generalizability coefficients, and indexes of dependability.

$p^{\bullet} \times i^{\circ} \times r^{\bullet}$ design. For the second study design, $p^{\bullet} \times i^{\circ} \times r^{\bullet}$, two univariate analyses were conducted for the MC and FR sections separately, using GENOVA (Crick & Brennan, 1983). In this multivariate design, the only covariance term that needed to be estimated was the

person facet, which was obtained by hand calculation. To be specific, for the $p^{\bullet} \times i^{\circ}$ multivariate design where each fixed level shares persons in common, but not items,

$$\hat{\sigma}_{MF}(p) = S_{MF}(p) = \frac{n_p}{n_p - 1}\left(\frac{\sum_p \bar{X}_{pM}\bar{X}_{pF}}{n_p} - \bar{X}_M\bar{X}_F\right),$$

where $\hat{\sigma}_{MF}(p)$ represents the estimated covariance between section universe scores, $S_{MF}(p)$ indicates the observed covariance between MC and FR section scores, n_p is the number of persons, $\bar{X}_{pM}$ and $\bar{X}_{pF}$ are person's MC and FR section mean scores, respectively, and $\bar{X}_M$ and $\bar{X}_F$ are the same as previously defined.

Note that assuming uncorrelated errors across sections, the errors for composite scores are the weighted sum of errors associated with each section. Also, the universe scores for the composites are the weighted sum of item-format specific universe scores plus two times the weighted covariance between sections. Relative error variance, absolute error variance, and universe score variance for the composite scores can be presented as

$$\sigma_C^2(\delta) = w_M^2\left[\frac{\sigma_M^2(pi)}{n_{iM}}\right] + w_F^2\left[\frac{\sigma_F^2(pi)}{n_{iF}} + \frac{\sigma_F^2(pr)}{n_{rF}} + \frac{\sigma_F^2(pir)}{n_{iF}n_{rF}}\right],$$

$$\sigma_C^2(\Delta) = w_M^2\left[\frac{\sigma_M^2(i)}{n_{iM}} + \frac{\sigma_M^2(pi)}{n_{iM}}\right] + w_F^2\left[\frac{\sigma_F^2(i)}{n_{iF}} + \frac{\sigma_F^2(r)}{n_{rF}} + \frac{\sigma_F^2(ir)}{n_{iF}n_{rF}} + \frac{\sigma_F^2(pi)}{n_{iF}} + \frac{\sigma_F^2(pr)}{n_{rF}} + \frac{\sigma_F^2(pir)}{n_{iF}n_{rF}}\right],$$

and

$$\sigma_C^2(p) = w_M^2\sigma_M^2(p) + w_F^2\sigma_F^2(p) + 2w_Mw_F\sigma_{MF}(p),$$

where w_M^2 = 65 and w_F^2 = 13.

All the facets including persons, items, and raters were regarded as random facets in this design.

Due to the way data were collected, the intact data from the special study could not be used with either a crossed or nested design. Thus, the special study data were divided into 8 folders and these folders were recombined in a way that subsets of the data could be analyzed as either a crossed or nested G study design. For example, referring to Figure 2, when folder1 and folder2 were combined, the data structure of this subset is a $p \times i \times r$ design with two items, two raters, and 55 persons. The same approach was applied to the remaining folders. As a result, four subsets were obtained, which were considered to be pseudo replications. Note that the replications cannot be viewed as entirely random since some replications involved the same persons. D study statistics were computed using the averaged G study variance components across the four subsets, with the expectation that they would serve as more accurate estimates

than the estimates from a single subset. The operational responses to the FR items (② in Figure 1) could not be used in this analysis because rater information was not available for the operational data. In summary, combined data were created using MC items from the operational data (① in Figure 1) and two ratings from the special study (③ and ④), but not the operational FR scores (②) due to the lack of rater information. Note that negative variance component estimates were retained for the pseudo replications, but average variance components (over replications) that were negative were set to zero.

$\boldsymbol{p^{\bullet} \times (r^{\circ}: i^{\circ})}$ **design.** An approach similar to the crossed design was employed for analyzing the data with this design. Data were obtained by combining the responses to the MC items from the operational data (① in Figure 1) and two ratings from the special study for the FR items (③ and ④). The only difference is that the $p \times (r: i)$ design was used for the FR section in the nested design. Therefore, for this multivariate design, the special study data were divided and recombined in order to create subsets of the data that could be analyzed as a nested G study design. For instance, combining folder1 and folder3 in Figure 2 would result in a data structure of a $p \times (r: i)$ design in which each item is scored by a different set of raters.

Issues of Interest

Number of raters. Operationally, scores for FR items are rated by a single rater. Thus, it is important to investigate whether reliability based on a single rater is satisfactory. To examine whether a single scoring still leads to an acceptable level of reliability, results based on one versus two raters were compared.

Number of items. One of the most critical factors influencing reliability is test length, or the number of items. Thus, tests with five different numbers of items were considered. In the operational setting, the AP German Language exam is administered in a three-hour session: 1 hour and 35 minutes for the MC section, and 1 hour and 25 minutes for the FR section. Therefore, for D study considerations, the pairs of the numbers of items per each section were chosen so that the administration time could be held as constant as possible. A similar approach was taken by Powers and Brennan (2009). Based on the time constraint, removing one FR item was worth adding 15 MC items. The five D studies considered were: 35 MC items and 6 FR items, 50 MC items and 5 FR items, 65 MC items and 4 FR items, 80 MC items and 3 FR items, and 95 MC items and 2 FR items.

However, it is important to note that this approach implicitly treats operational FR items as randomly parallel (i.e., FR items are regarded as interchangeable in the universe of items), which may not be completely justifiable. Specifically, there are some different characteristics among the four operational FR items in terms of skills (or content) that each item is intended to measure. Each of the four items is designed to measure different skills including interpersonal writing, presentational writing, interpersonal speaking, and presentational speaking.

Results

Results are presented for each of the three multivariate G theory designs. For each study design, the G and D study results are provided first followed by a discussion concerning the impact of changing the number of raters and items. Then, a detailed comparison of the three designs is presented.

Correlations between the MC and FR section scores were computed to examine the degree of dimensionality. The observed correlations between MC (① in Figure 1) and FR section scores for these studies (②, ③ and ④) were .8390, .8591, and .8299 for the operational dataset and two special study datasets, respectively. The disattenuated correlations were .9352, .9658, and .9336 for the three datasets.

$p^{\bullet} \times i^{\circ}$ Design

Table 1 provides the estimated G study variance components based on the multivariate $p^{\bullet} \times i^{\circ}$ design. In Table 1, the first three columns show the estimated variance components using various sources of data including the operational dataset and the two special study datasets. The last column presents the average values over those three estimates. As mentioned previously, these average values were treated as final estimates and used in computing D study statistics for the $p^{\bullet} \times I^{\circ}$ design.

As shown in Table 1, the variance components associated with the FR section are always larger than those for the MC section. This is almost certainly mainly attributable to the fact that the maximum possible score for each FR item is 5 times larger than that for MC items. In terms of variance components, $\hat{\sigma}^2(i)$ is small regardless of section, implying that variability attributable to items is not substantial. Also, for the FR section, the variability due to persons, $\hat{\sigma}^2(p)$, is larger than $\hat{\sigma}^2(pi)$, where the latter indicates variability resulting from different rank ordering of persons by item plus all other unidentified residual errors. For the MC section,

however, the reverse pattern is observed. The estimated covariance component, $\hat{\sigma}_{mf}$, is similar across the three datasets.

Table 2 presents D-study results for composite scores, including universe score variance, error variances, and reliability estimates based on various pairs of numbers of items for the two sections. All computation was done using the mean score metric, following the usual G theory convention. The operational condition is marked with an asterisk (*). The first row contains the universe score variance for the composite score, $\hat{\sigma}_C^2(\tau)$. The second row shows the estimated relative error variance $\hat{\sigma}_C^2(\delta)$, followed by the absolute error variance $\hat{\sigma}_C^2(\Delta)$. The last two rows present estimated generalizability and dependability coefficients. As expected, relative error variance is consistently smaller than absolute error variance, and consequently, the generalizability coefficient is larger than the index of dependability.

Number of raters. For the $p^{\bullet} \times i^{\circ}$ design, the effects of changing the number of raters were not estimable because the errors attributable to raters were completely confounded with the errors due to items

Number of items. Regarding the number of items, keeping the administration time as a constraint, the operational condition (65MC and 4FR) seems to provide reasonable results in terms of reliability estimates, as shown in Figure 2. In fact, adding 1 FR item and removing 15 MC items from the operational condition (i.e., 50 MC and 5 FR) results in the largest reliability estimates among the five conditions. A study conducted by Powers and Brennan (2009) demonstrated the highest composite score reliability for both AP Biology and AP World History exams under operational conditions, which is inconsistent with the findings of the current study. However, the magnitude of the differences in composite score reliability among the five pairs of numbers of items does not seem substantial, since the maximum difference is approximately .035.

$\boldsymbol{p^{\bullet} \times i^{\circ} \times r^{\bullet}}$ Design

As previously discussed, two separate analyses were conducted for each of the MC and FR sections, using a different design for each section. To obtain reliability estimates for the composite scores, the first step was to estimate the variance components for each section using an item-format specific univariate design. The variance component estimates for the MC section are provided in Table 3.

The variance component estimates for the FR section are provided in Table 4. As mentioned previously, four pseudo replications were identified for analyzing the special study data. Table 4 includes the numbers of folders that were combined to achieve the crossed design and the variance components estimated from each replication. The average estimated variance components and the standard deviations of the estimated variance components (over four pseudo replications) are reported in the last two columns of Table 4. Again, the average variance components were used to obtain results for subsequent D studies. The standard deviations serve as estimated empirical standard errors (SE) for estimated variance components. The estimated standard error for $\hat{\sigma}^2(p)$ is almost one-third of the average; the estimated standard error for $\hat{\sigma}^2(i)$ is almost one-half of the average, and so on. These somewhat large estimated SEs are likely due to small sample sizes.

The G study variance components for the $p^{\bullet} \times i^{\circ} \times r^{\bullet}$ design are in Table 5. The variance components for each section were obtained directly from each univariate G study analysis. Thus, their values are the same as those presented in Table 3 for the MC section and Table 4 for the FR section. The estimated universe score covariances between the MC and FR sections are also provided in Table 5. D study results for the composite scores are provided in Table 6.

Number of raters. Table 6 suggests, as expected, that employing two raters leads to higher reliability estimates than using a single rater for both types of reliability. Greater improvement in reliability for the smaller number of FR items could be achieved by using two raters. As a result, the maximum difference between single versus double raters is found for the 95 MC and 2 FR items condition. This finding is reasonable given that the number of raters affects only the errors associated with FR items. So, by adding more FR items, the number of raters plays a less important role in decreasing the errors associated with FR items.

However, it is important to notice that even the maximum difference between single and double raters is nearly .02 in magnitude. The reason that increasing the number of raters does not contribute to increasing reliability estimates is a direct result of the G study estimated variance components in Table 4; i.e., the estimated variance components associated with the rater effect (*r*, *pr*, or *ir*) are all close to zero.

Number of items. Regarding the number of items, a similar pattern is found as the $p^{\bullet} \times i^{\circ}$ design. The 50 MC and 5 FR condition provides the largest reliability estimates, although the

differences between the five conditions are not very large. As mentioned before, the differences become larger for the single rater than for the double raters condition.

$p^{\bullet} \times (r^{\circ}: i^{\circ})$ Design

An approach similar to the crossed design was taken for the nested design. The estimated variance components for the MC and FR sections can be found in Tables 3 and 7, respectively. Specifically, Table 7 provides variance components for the FR section for the four pseudo replications, their average values, and their standard deviations. In general, $\hat{\sigma}^2(i)$ and $\hat{\sigma}^2(r{:}\,i)$ are substantially smaller than $\hat{\sigma}^2(p)$ across all folders. Compared to the rater and item effects, $\hat{\sigma}^2(pi)$ is larger, suggesting that the rank ordering of persons varies by item. Also, the residual term, $\hat{\sigma}^2(pr{:}\,i)$, tends to be large, which is typically expected because it includes all the variability that is not defined in the design. The empirical standard errors are relatively large because of small sample sizes.

Table 8 shows the G study variance components for the $p^{\bullet} \times (r^{\circ}: i^{\circ})$ design. Again, the variance components for each section came directly from the corresponding univariate G study results. Additionally, the universe score covariances between the MC and FR sections were computed under the nested design. Table 9 provides D study statistics based on the average G study variance components shown in the far right column of Table 8.

Number of raters. Regarding the number of raters, patterns that are very similar to the crossed design are observed. Overall, two raters produce higher reliability estimates than a single rater, and the differences in the magnitude of reliability estimates across various D-study sample sizes are not substantial.

Number of items. As with the previous two designs, $p^{\bullet} \times i^{\circ}$ and $p^{\bullet} \times i^{\circ} \times r^{\circ}$, the largest reliability estimates are found with the 50 MC and 5 FR condition. The operational condition (i.e., 4 FR and 65 MC) provides the second largest reliability estimates. The tendency among the five conditions is very similar to the other two designs.

Comparison of Designs

The generalizability coefficients estimated from the series of D studies are displayed in Figure 2. Figure 3 depicts the index of dependability. The red and purple lines show the estimates based on double raters for crossed and nested designs, respectively. The estimates based on a single rater are represented with green and bright blue lines.

Among the three multivariate designs, the crossed design tends to provide the highest reliability estimates across all study conditions. Usually, however, all other things being equal, the nested design typically leads to smaller error variances and, in turn, larger reliability coefficients (Brennan, 2001). The reverse pattern is found in this study, however, largely because these two designs were associated with different datasets. Although the same datasets were used for the crossed and nested design analyses, through the data splitting process, datasets were manipulated in a different way. Another plausible explanation is that the estimated universe score variance is larger for the crossed design than the nested design, as presented in Tables 6 and 9. Note also that although the sample size for the initial dataset was not terribly small, the sample sizes for the datasets that were actually analyzed became small due to the data splitting process. Small sample sizes resulted in negative variance component estimates for the rater-related effects. The sampling variability appears to have contributed to this unusual observation.

A comparison of the designs with and without the rater facet (i.e., $p^{\bullet} \times i^{\circ} \times r^{\circ}$and $p^{\bullet} \times (r^{\circ}: i^{\circ})$ versus $p^{\bullet} \times i^{\circ}$) indicates that reliability coefficients tend to be underestimated when; (a) the rater effect is ignored, and (b) the rater facet is considered as random in the investigator's universe of generalization. In the operational setting considered in this paper, it is impossible to estimate the error variance resulting from rater variability because only one rater is used. Thus, the rater facet is hidden and (effectively) fixed in this context. Brennan (2001) notes that estimates of reliability tend to be higher than they should be when a facet is hidden and fixed. The results based on the AP German Language exam, as shown in Figures 2 and 3, suggest that composite score reliability might be slightly underestimated with the conventional estimation procedure, which does not take rater variability into account.

Summary and Discussion

Estimating reliability for mixed-format tests is less straightforward than estimating reliability for a test consisting of a single item format. The primary benefit of using G theory comes from its flexibility. G theory provides a useful psychometric tool even for complex data structure. This study illustrated with real data analyses several multivariate G study designs that can be used for mixed-format tests. One of the strengths of this study comes from the use of data obtained from a special study, which is an approach that has not yet been much discussed or used in the literature. Multiple sources of data – operational data and special study data – allow for an effective examination of rater effects. In addition, use of a multivariate design involving different

random models for different levels of the fixed facet (i.e., MC vs. FR sections) suggests a potential framework for analyzing and interpreting the psychometric properties of mixed-format tests.

The study results showed that error variance and reliability estimates can be biased when rater effects are ignored. This finding may have practical implications. The composite score reliability for AP exams is typically obtained by combining the reliabilities estimated from each section. The FR section reliability is typically computed ignoring rater variability. All other things being equal, this is likely to lead to inflated estimates of the composite score reliability. To report more accurate reliability estimates, all possible sources of errors need to be specified in the estimation procedure.

It was also found that the crossed rater facet produced higher reliability estimates than the nested rater facet, although a nested design usually provides higher reliability than the crossed design. When choosing a more suitable design, the decision will mainly depend on investigator's conceptual interpretation of the available data structure vis-à-vis the universe of admissible observations. Additionally, it is worth stressing that estimates are subject to sampling variability, so no definitive statement regarding study design superiority should be made based solely on the results of this study, because the differences in estimates between the two designs were negligible.

Regarding the number of raters, doubling the number of raters did not seem to improve reliability appreciably, which supports the operational use of single scoring. In terms of the number of items per section, the operational condition provided satisfactorily high reliability coefficients. Note also that eliminating 15 MC items and adding 1 FR item to the operational condition led to only slight increases in estimates.

Limitations and Future Considerations

Some limitations in the present study should be noted. First, sample sizes were small. As discussed previously, negative variance components were found due to small sample sizes. Meanwhile, it has been reported that, with well-trained raters, the rater facet usually does not contribute much to the variability in observed scores (Brennan & Johnson, 1995; Brennan, 2000). The presence of negative estimates usually does not have substantial impact on the study results. Nonetheless, as is true for any study, a larger sample size would have improved the estimation precision to some extent. Second, only one AP exam was used. Moses and Kim

(2015) pointed out that some psychometric characteristics of a test, such as the correlation between MC and FR section scores or correlation between FR scores obtained from different raters, could influence D study results and interpretation. Future research could further investigate the impact of these factors using multiple datasets with various test characteristics. Note that the multivariate analyses employed in this study required data obtained from multiple raters, which is often not available from the operational data collection design.

References

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R. L. (2001a). *Generalizability theory.* New York, NY: Springer-Verlag.

Brennan, R. L. (2001b). *mGENOVA* [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L. (2016a). *Nominal weights in multivariate generalizability theory*. (CASMA Research Report No. 50). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L. (2016b). *Using G Theory to Examine Confounded Effects: "The Problem of One."* (CASMA Research Report No. 51). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on https://education.uiowa.edu/centers/casma)

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*, 9-12.

Clauser, B. E., Balog, K., Harik, P., Mee, J., & Kahraman, N. (2009). A multivariate analysis of history-taking and physical examination scores from the USMLE Step 2 Clinical Skills Examination. *Academic Medicine, 84*, 586-589.

Crick, J. E., & Brennan, R. L. (1983). *GENOVA* [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*, 191-203.

Jarjoura, D., Early, L., & Androulakakis, V. (2004). A multivariate generalizability model for clinical skills assessments. *Educational and Psychological Measurement, 64*, 22-39.

Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. *Applied Psychological Measurement, 39*, 314-329.

Powers, S., & Brennan, R.L. (2009). *Multivariate Generalizability Analyses of Mixed-Format Advanced Placement Exams.* (Research Report No. 29). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Table 1

G Study Variance and Covariance Components for the $p^{\bullet} \times i^{\circ}$ Design

	Operational Data		Special Study Data1		Special Study Data2		Average	
	MC	FR	MC	FR	MC	FR	MC	FR
$\hat{\sigma}^2(p)$	.03156	1.12382	.03156	1.09841	.03156	1.02308	.03156	1.08177
$\hat{\sigma}_{mf}(p)$	.17561		.17929		.16726		.17405	
$\hat{\sigma}^2(i)$	.02630	.05328	.02630	.08278	.02630	-.00011	.02630	.04532
$\hat{\sigma}^2(pi)$	.17479	.62104	.17479	.69245	.17479	.65149	.17469	.62499

Table 2

Composite Error Variance and Coefficients for the $p^{\bullet} \times I^{\circ}$ Design

	D Studies				
n'_{iM}	35	50	65*	80	95
n'_{iF}	6	5	4*	3	2
$\hat{\sigma}^2_C(\tau)$	610.3070	610.3070	610.3070	610.3070	610.3070
$\hat{\sigma}^2_C(\delta)$	39.5481	36.9082	39.0345	46.1287	63.1201
$\hat{\sigma}^2_C(\Delta)$	44.0002	40.6633	42.6600	50.0725	68.1220
$E\hat{\rho}^2$	.9391	.9430	.9399	.9297	.9063
$\hat{\Phi}$	.9328	.9375	.9347	.9242	.8996

Note. Asterisk (*) indicates the operational condition.

Table 3

G Study Variance Components for the $p \times i$ Design (MC Items only)

	G Study
$\hat{\sigma}^2(p\|r)$	.0316
$\hat{\sigma}^2(i\|r)$	.0263
$\hat{\sigma}^2(pi\|r)$	.1748

Table 4

G Study Variance Components for the $p \times i \times r$ Design (FR Items only)

	Folder1/2	Folder3/4	Folder5/8	Folder6/7	Average	SD
$\hat{\sigma}^2(p)$	.9955	1.7508	1.2311	.6191	1.1491	.4103
$\hat{\sigma}^2(i)$	.0429	.0082	.0269	.0835	.0404	.0278
$\hat{\sigma}^2(r)$	.0025	.0370	-.0940	.0019	-.0131	.0488
$\hat{\sigma}^2(pi)$	.0934	.2736	.4361	.2406	.2609	.1218
$\hat{\sigma}^2(pr)$	-.0480	-.0007	-.0356	.0814	-.0007	.0505
$\hat{\sigma}^2(ir)$	-.0042	-.0059	.1822	.0016	.0434	.0802
$\hat{\sigma}^2(pir)$	.2724	.3423	.3872	.2114	.3033	.0670

Table 5

G Study Variance and Covariance Components for the $p^{\bullet} \times i^{\circ} \times r^{\circ}$ Design

	Folder1/2		Folder3/4		Folder5/8		Folder6/7		Average	
	MC	FR	MC	FR	MC	FR	MC	FR	MC	FR
$\hat{\sigma}^2(p)$	.0316	.9955	.0316	1.7508	.0316	1.2311	.0316	.6191	.0316	1.1491
$\hat{\sigma}_{mf}(p)$	.1778		.2351		.1619		.1185		.1733	
$\hat{\sigma}^2(i)$	.0263	.0429	.0263	.0082	.0263	.0269	.0263	.0835	.0263	.0404
$\hat{\sigma}^2(r)$	-	.0025	-	.0370	-	-.0940	-	.0019	-	-.0131
$\hat{\sigma}^2(pi)$	.1748	.0934	.1748	.2736	.1748	.4361	.1748	.2406	.1748	.2609
$\hat{\sigma}^2(pr)$	-	-.0480	-	-.0007	-	-.0356	-	.0814	-	-.0007
$\hat{\sigma}^2(ir)$	-	-.0042	-	-.0059	-	.1822	-	.0016	-	.0434
$\hat{\sigma}^2(pir)$	-	.2724	-	.3423	-	.3872	-	.2114	-	.3033

Table 6

Composite Error Variance and Coefficients for the $p^{\bullet} \times I^{\circ} \times R^{\circ}$ Design

	D Studies									
n'_{iM}	35	50	65	80	95	35	50	65*	80	95
n'_{iF}	6	5	4	3	2	6	5	4*	3	2
n'_{r}	2	2	2	2	2	1	1	1*	1	1
$\hat{\sigma}^2_C(\tau)$	620.4159	620. 4159	620. 4159	620. 4159	620. 4159	620.4159	620. 4159	620. 4159	620. 4159	620. 4159
$\hat{\sigma}^2_C(\delta)$	32.7210	28.7148	28.7922	32.4719	42.6345	36.9925	33.8406	35.1995	41.0149	55.4489
$\hat{\sigma}^2_C(\Delta)$	37.6450	33.0361	33.1255	37.3592	49.0516	42.5276	38.8954	40.4495	47.1246	63.6997
$E\hat{\rho}^2$	.9499	.9558	.9557	.9503	.9357	.9437	.9483	.9463	.9380	.9180
$\hat{\Phi}$	.9428	.9494	.9493	.9432	.9267	.9359	.9410	.9388	.9294	.9069

Note. Asterisk (*) indicates the operational condition.

Table 7

G Study Variance Components for $p \times (r{:}\, i)$ Design (FR Items only)

	Folder1/3	Folder2/4	Folder5/6	Folder7/8	Average	SD
$\hat{\sigma}^2(p)$	.9673	1.6562	.7975	.8616	1.0707	.3435
$\hat{\sigma}^2(i)$	.1026	-.0152	-.0016	.2172	.0758	.0935
$\hat{\sigma}^2(r{:}\, i)$	.0150	.0145	.0398	.0519	.0303	.0161
$\hat{\sigma}^2(pi)$	.2383	.2515	.0756	.7921	.3394	.2704
$\hat{\sigma}^2(pr{:}\, i)$	.1805	.3855	.3166	.3277	.3026	.0752

Table 8

G Study Variance and Covariance Components for the $p^{\bullet} \times (r^{\circ}: i^{\circ})$ Design

	Folder1/3		Folder2/4		Folder5/6		Folder7/8		Average	
	MC	FR	MC	FR	MC	FR	MC	FR	MC	FR
$\hat{\sigma}^2(p)$	.0316	.9673	.0316	1.6562	.0316	.7975	.0316	.8616	.0316	1.0707
$\hat{\sigma}_{mf}(p)$	.1859		.2271		.1337		.1467		.1733	
$\hat{\sigma}^2(i)$	.0263	.1026	.0263	-.0152	.0263	-.0016	.0263	.2172	.0263	.0758
$\hat{\sigma}^2(r:i)$	-	.0150	-	.0145	-	.0398	-	.0519	-	.0303
$\hat{\sigma}^2(pi)$	.1748	.2383	.1748	.2515	.1748	.0756	.1748	.7921	.1748	.3394
$\hat{\sigma}^2(pr:i)$	-	.1805	-	.3855	-	.3166	-	.3277	-	.3026

Table 9

Composite Error Variance and Coefficients for the $p^{\bullet} \times (R^{\circ}: I^{\circ})$ Design

	D Studies									
n'_{iM}	35	50	65	80	95	35	50	65	80	95
n'_{iF}	6	5	4	3	2	6	5	4	3	2
n'_r	2	2	2	2	2	1	1	1	1	1
$\hat{\sigma}^2_C(\tau)$	607.1663	607.1663	607.1663	607.1663	607.1663	607.1663	607.1663	607.1663	607.1663	607.1663
$\hat{\sigma}^2_C(\delta)$	34.9222	31.3563	32.0941	36.8744	49.2382	39.1839	36.4702	38.4865	45.3976	62.0230
$\hat{\sigma}^2_C(\Delta)$	40.6588	36.6527	37.6462	43.3869	58.0931	45.3471	42.2787	44.6787	52.7636	72.1581
$E\hat{\rho}^2$	.9456	.9509	.9498	.9427	.9250	.9394	.9433	.9404	.9304	.9073
$\hat{\Phi}$	.9372	.9431	.9416	.9333	.9127	.9305	.9349	.9315	.9200	.8938

	Operational Data	Special Study Data1	Special Study Data2
MC	MC1 . . . MC65 ①	NA	NA
FR	FR1 . . . FR4 ②	FR1 . . . FR4 ③	FR1 . . . FR4 ④

Figure 1. Data structure for three datasets.

	I1		I2		I3		I4	
P1	R1	R2	R1	R2	R3	R4	R3	R4
. .	. Folder1 .		. Folder2 .		. Folder3 .		. Folder4 .	
P55	R1	R2	R1	R2	R3	R4	R3	R4
P56	R5	R6	R7	R8	R7	R8	R5	R6
. .	. Folder5 .		. Folder6 .		. Folder7 .		. Folder8 .	
P109	R5	R6	R7	R8	R7	R8	R5	R6

Figure 2. Data structure for the special study (FR items only).

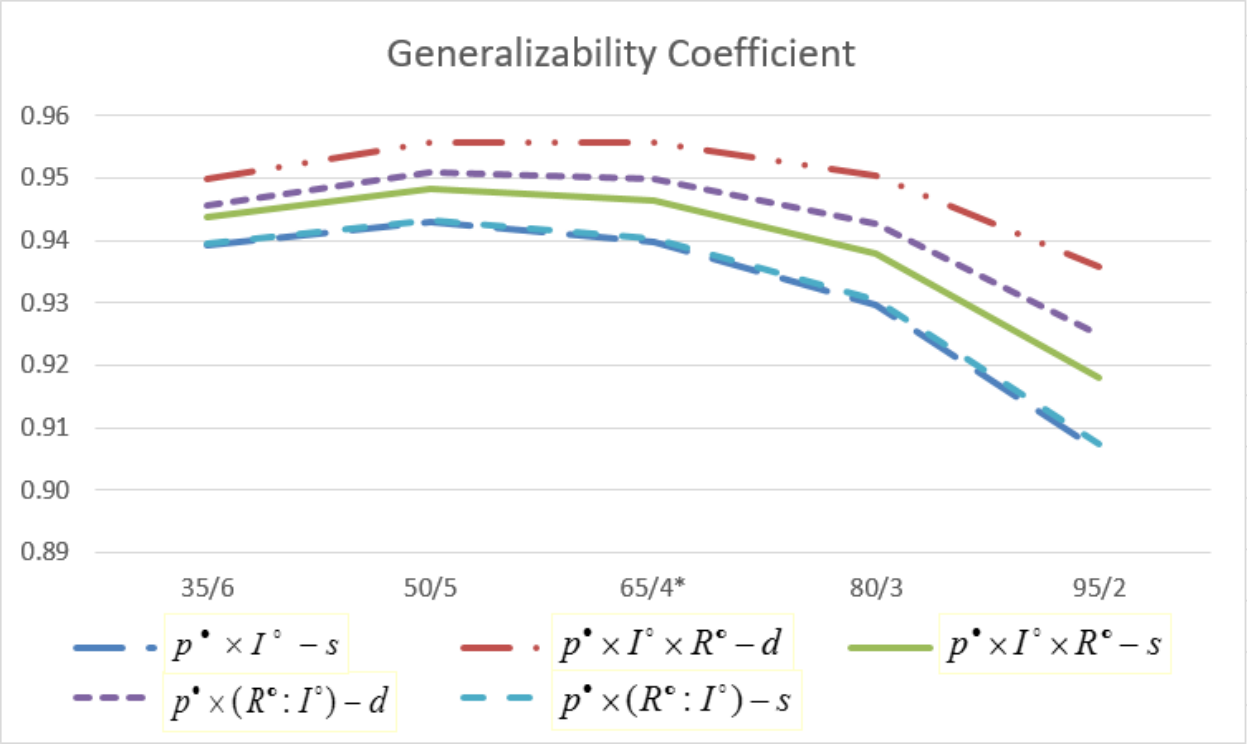

Figure 2. Generalizability coefficients for various numbers of raters and items.
Note. Asterisk (*) indicates the operational condition.
-d represents double raters and *–s* represents a single rater.

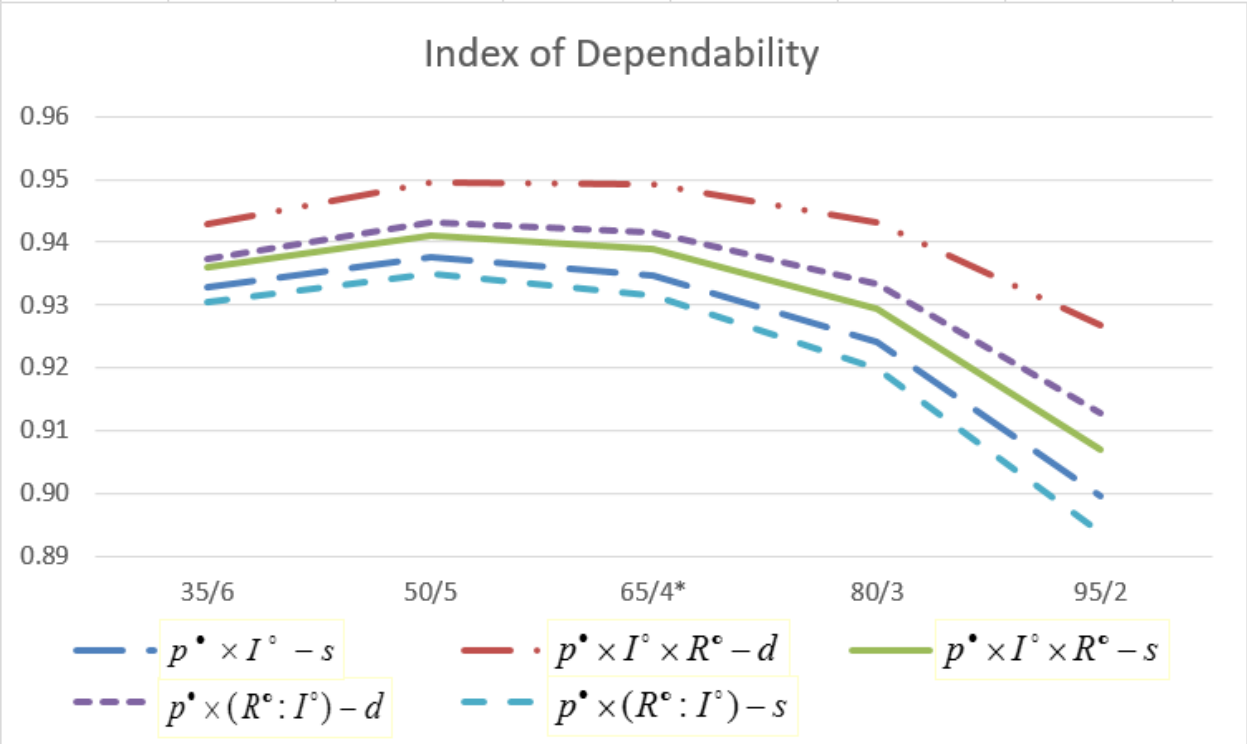

Figure 3. Index of dependability for various numbers of raters and items.
Note. Asterisk (*) indicates the operational condition.
-d represents double raters and *–s* represents a single rater.

Chapter 8: Evaluation of Scale Transformation Methods with Stabilized Conditional Standard Errors of Measurement for Mixed-Format Tests

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

Abstract

This study focuses on evaluating three raw-to-scale score transformation methods that can produce stabilized conditional standard errors of measurement (CSEMs). The transformation methods considered are the arcsine transformation and two recently developed methods, the general variance stabilization method and the cubic transformation method. Mixed-format tests that consist of dichotomous and polytomous items are the focus of the evaluation, although tests with only dichotomously scored items or polytomously scored items are also considered. The IRT methods for estimating CSEMs for mixed-format tests are reviewed. The three scale score transformation methods are applied to five pseudo-tests constructed based on an operational test to examine their performance in terms of stabilizing CSEMs. The results suggest that, in general, all three transformation methods were effective at stabilizing CSEMs for a majority of the score points in the middle of the score range. The arcsine transformation method was affected by the test composition, yielding more stable CSEMs for mixed-format tests with a larger portion of dichotomous items (at least 50%).

Evaluation of Scale Transformation Methods with Stabilized Conditional Standard Errors of Measurement for Mixed-Format Tests

Conditional standard errors of measurement (CSEMs), which are estimates of the standard errors conditional on different score levels, are important indices of measurement accuracy. CSEMs can assist test users in the proper interpretation of reported scores. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) suggest that CSEMs should be provided in units of the reported scores, and if CSEMs are not constant across score levels, they should be reported at several levels.

As suggested by previous studies, the CSEMs of raw scores typically vary along the score scale (Feldt & Brennan, 1989; Lord, 1955; Mollenkopf, 1949). However, when raw scores are transformed to scale scores for reporting purposes, nonlinear scale score transformations can result in a very different pattern of scale-score CSEMs than the pattern observed for raw scores (Kolen & Brennan, 2014). To avoid the confusion caused by multiple reported CSEMs, it is very useful to simplify score interpretation by employing a scale score transformation that produces approximately equal CSEMs along the score scale, because then only one value for the CSEMs needs to be reported.

The arcsine transformation suggested by Kolen (1998) is the most well-known scaling method in the psychometric literature for stabilizing the magnitude of the CSEMs. Recently, two new scale score transformation methods, the general variance stabilization (GVS) transformation (Li, Woodruff, Thompson, & Wang, 2014) and the cubic transformation (Moses & Kim, 2016), were developed for the same purpose. Li et al. (2014) stated that the GVS transformation is a more general and straightforward method compared to the arcsine transformation because it can be used with various psychometric models such as true score models with binomial error and the unidimensional and multidimensional IRT models. Moses and Kim (2016) proposed and compared the cubic transformation method with the arcsine and GVS transformation methods. They concluded the cubic transformation produced smoother and less extreme scale scores and scale-score CSEMs than the other two transformations. Another advantage of the cubic method was that the resulting scale score distributions were more nearly symmetric than were the other two methods for stabilizing CSEMs. It should be noted that Moses and Kim's study only considered tests containing dichotomously scored items. It is still unclear how well these results will generalize to mixed-format tests that contain both dichotomously and polytomously scored

items. The extension to mixed-format tests complicates the evaluation of these three transformation methods. Polytomously scored items, such as free-response (FR) items, typically yield scores that are less reliable than dichotomously scored items, such as multiple-choice (MC) items, per unit of testing time. Therefore, the effectiveness of the three scale transformation methods on stabilizing CSEMs may vary by how each item type contributes to the total points or the sparseness of the raw score distribution.

The purpose of this study was to evaluate scale transformation methods that produce stabilized CSEMs including the arcsine, GVS, and cubic transformation methods for mixed-format tests. Mixed-format tests that consist of dichotomously and polytomously scored items were the focus of the evaluation, but tests with only dichotomously scored MC items or only polytomously scored FR items were also considered. IRT methods were used to estimate the scale-score CSEMs. More specifically, the following research questions were addressed:

1. For mixed-format tests, which scale transformation method among the arcsine, GVS, and cubic transformations yields the most stable CSEMs along the score scale? Are the results consistent with the tests containing only dichotomously scored items?
2. Does the composition of the mixed-format test in terms of MC and FR item scores affect the three scale transformation methods in terms of producing stable CSEMs? If so, to what extent?

Method

Data Source and Construction of Pseudo-Tests

This study uses the College Board Advanced Placement Program® (AP®) Chemistry exam administered in 2013. This exam includes one dichotomously scored MC section and one polytomously scored FR section. Real test data collected from more than 20,000 examinees were included. Five psuedo tests were created based on the original AP Chemistry exam. The number of items, possible raw score points for each section for the original AP Chemistry, and pseudo-tests are summarized in Table 1. Test 1 contains only MC items, and Test 5 contains only FR items. Tests 2 to 4 are constructed as mixed-format tests that consist of a set of MC items and a set of FR items. The percentage of points for the MC section for Tests 2 to 4 are 75%, 51%, and 25%, respectively. The total possible raw score points across tests range from 61 to 81. Table 2 displays descriptive statistics and reliabilities for the five psuedo-tests. Reliabilities were computed based on the IRT method, which is discussed later. The means of the four tests vary

from 24.71 to 43.57, and the standard deviations range from 13.74 to 17.97. The reliabilities for the five tests range from 0.93 to 0.95. The observed raw score distributions for the five tests are shown in Figure 1. As can be seen, the observed raw score distributions for Tests 1 and 2 are approximately symmetrical, and the distributions for Tests 3 to 5 are skewed to the right.

Item Calibration

The computer program flexMIRT (Cai, 2015) was used to estimate all the item parameters. The calibration run successfully converged. Forty nine quadrature points evenly selected from -6.0 to 6.0 were used for marginal maximum likelihood estimation. Dichotomously scored MC items were fit using the three-parameter logistic (3PL; Lord, 1980) model and the polytomously scored FR items were calibrated using Muraki's generalized partial credit (GPC; Muraki, 1992) model. The two models are briefly described here. With the 3PL model, the probability of answering an item correct for a person with ability of θ is

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[1.7a_i(\theta - b_i)]}{1 + \exp[1.7a_i(\theta - b_i)]}, \quad (1)$$

where a_i, b_i and c_i are item *i*'s discrimination, difficulty, and pseudo-guessing parameters, respectively. Under the GPC model, each item has a discrimination and overall difficulty parameters, and each item category also has its own difficulty parameter. The GPC model is mathematically expressed as

$$P_{ik}(\theta) = P_{ik}\left(\theta; a_i, b_i, d_{i2}, \ldots\ d_{im_i}\right) = \frac{\exp[\sum_{h=1}^{k} 1.7a_i(\theta - b_i + d_{ih})]}{\sum_{g=1}^{m_i}\exp[\sum_{h=1}^{g} 1.7a_i(\theta - b_i + d_{ih})]}, \quad (2)$$

where $P_{ik}(\theta)$ is the probability of a person with ability of θ correctly answering item *i* with category parameters *k*, d_{ih} is the item category difficulty parameter, and m_i is the number of categories for item *i*.

Procedures for Estimating CSEMs and Reliability for Scale Scores

For tests containing both dichotomous and polytomous items, the conditional probability distribution of raw scores is obtained by an extension of the Lord-Wingersky algorithm provided by Hanson (1994), Thissen, Pommerich, Billeaud, and Williams (1995), and Wang, Kolen, and Harris (2000). Assume a test consists of a total of *K* items (including both dichotomous and polytomous items) with maximum total raw score of *N*. Let U_i be a random variable for the score on item *i* with maximum score n_i, and u_i be a raw score of item *i* (u_i = 0, 1, …, n_i). Let *Y* be a random variable of raw score for an examinee with ability θ, and *y* be the total raw score for that

examinee (y = 0, 1, …, N). Thus, the recursive algorithm for the conditional distribution of the total raw score given ability $P(Y = y|\theta)$ can be expressed as follows:

For item i = 1,

$$P(Y_1 = y|\theta) = P(U_1 = y|\theta), \text{ for } y = 0,1, \dots, n_1. \quad (3)$$

For item i = 2, …, K,

$$P(Y_i = y|\theta) = \sum_{u_i=0}^{n_i} P(Y_{i-1} = y - u_i|\theta)\, P(U_i = u_i|\theta), \text{ for } y = 0,1, \dots, N. \quad (4)$$

The conditional probability distribution of raw scores is found by repeating Equations 3 and 4 until all K items are included in the recursive procedure.

The IRT approach for estimating CSEMs and reliability for scale scores has been illustrated by many researchers in the literature (Ban & Lee, 2007; Kolen & Lee, 2011; Kolen, Zeng, & Hanson, 1996; Kolen, Wang, & Lee, 2012; Wang et al., 2000). Let $sc(y)$ present a non-linear raw-to-scale score transformation function. The conditional distribution of scale scores given θ can be generated through the conditional distribution of raw scores given θ obtained using Equations 3 and 4. Thus, the true scale score at θ, or the mean of the conditional scale score distribution at θ is

$$\mu_{sc(y)|\theta} = \xi_{sc} = \sum_{y=0}^{N} sc(y)P(Y = y|\theta). \quad (5)$$

The conditional standard error of measurement for scale scores at θ is

$$\sigma_{sc(y)|\theta} = \sqrt{\sum_{y=0}^{N} \left[sc(y) - \mu_{sc(y)|\theta}\right]^2 P(Y = y|\theta)}. \quad (6)$$

The overall error variance for raw scores can be expressed as

$$\sigma^2[E] = \sum_{\theta} \sigma^2_{y|\theta}\psi(\theta)d\theta, \quad (7)$$

where $\psi(\theta)$ is the density function for ability in the population. A summation is used instead of an integral because a finite number of quadrature points are used. Similarly, the overall error variance for scale scores over the theta distribution is

$$\sigma^2\left[E_{sc(y)}\right] = \sum_{\theta} \sigma^2_{sc(y)|\theta}\psi(\theta)d\theta. \quad (8)$$

The marginal mean and standard deviation of scale scores for the population can be expressed as

$$\mu_{sc(y)} = \sum_{y=0}^{N} sc(y)P(Y = y), \tag{9}$$

and

$$\sigma_{sc(y)} = \sqrt{\sum_{y=0}^{N} \left[sc(y) - \mu_{sc(y)|\theta}\right]^2 P(Y = y)}, \tag{10}$$

where $P(Y = y)$ is the marginal distribution of raw scores. The reliability of scale scores is

$$rel_{sc(y)} = 1 - \frac{\sigma^2\left[E_{sc(y)}\right]}{\sigma^2{}_{sc(y)}}. \tag{11}$$

Scale Score Transformation Methods for CSEM Stabilization

Arcsine transformation. The arcsine transformation was proposed by Freeman and Tukey (1950) to stabilize standard deviations of binomial variates. The distribution of number-correct scores given true score is considered to be binomial or compound binomial in strong true score models (Lord, 1965) and in IRT. Therefore, the arcsine transformation can be used to stabilize error variance (Jarjoura, 1985; Wilcox, 1981). Kolen (1988) described the procedure for using the arcsine transformation to stabilize the CSEM of scale scores for tests consisting of dichotomous items. Ban and Lee (2007) extended the application of arcsine transformation to mixed-format tests. As suggested by Ban and Lee, the transformation of raw score *y* is

$$g(y) = g(y|N) = \frac{1}{2}\left(\sin^{-1}\sqrt{\frac{y}{N+1}} + \sin^{-1}\sqrt{\frac{y+1}{N+1}}\right), \tag{12}$$

where $\sin^{-1}$ is the arcsine function with its argument expressed in radians and *N* is the maximum raw score of the mixed-format test. To construct a scale with a desired mean μ_{sc} and constant standard error of measurement (SEM) sem_{sc}, the arcsine transformed score $g(y)$ can be linearly transformed using

$$sc_{Y,arcsine} = sc[g(y)] = \mu_{sc} + \frac{sem_{sc}}{sem_{g(y)}}\left[g(y) - \mu_{g(y)}\right], \tag{13}$$

where $\mu_{g(y)}$ and $sem_{g(y)}$ are the mean and estimated SEM for the arcsine transformed score, $g(y)$, and the mean and standard deviation can be computed using Equations 9 and 10.

GVS transformation. The GVS transformation was proposed by Li, Woodruff, Thompson, and Wang (2014) to equalize CSEMs. The delta method (Casella & Berger, 2002) suggests that

$$\sigma^2_{h(y)|\theta} \approx \left[\frac{\partial h(y)}{\partial y}\right]^2_{y=\xi} \cdot \sigma^2_{y|\theta}, \tag{14}$$

where $\sigma^2_{h(y)|\theta}$, and $\sigma^2_{y|\theta}$ are the conditional measurement error variance of GVS transformed scores and raw scores at θ, respectively. The subscript on the derivative represents that it is evaluated with y equal to its true raw score ξ. Let Equation 14 be equal to the square of the desired SEM sem_{sc}. The GVS transformed score $h(y)$ can be found by taking

$$h(y) = \int_{\theta_L}^{\theta} \frac{sem_{sc}}{\sigma_{y|\theta}} d\theta. \tag{15}$$

The integral in Equation 15 typically is evaluated using numerical integration methods, which rely on summations. Thus, the GVS transformed score $h(y)$ can be expressed as

$$h(y) = \sum_{h=1}^{y} \frac{sem_{sc}}{CSEM_h}; \tag{16}$$

that is, the GVS transformed score of y is a sum of the ratios of the desired constant SEM to the CSEM at raw scores 1 to y. The GVS transformed score of 0 is set to 0. Similar to the arcsine transformation, the GVS transformed score $h(y)$ can be linearly transformed to obtain a scale with desired mean μ_{sc}, and constant SEM sem_{sc}:

$$sc_{y,GVS} = sc[h(y)] = \mu_{sc} + \frac{sem_{sc}}{sem_{h(y)}}\left[h(y) - \mu_{h(y)}\right], \tag{17}$$

where $\mu_{h(y)}$ and $sem_{h(y)}$ are the mean and the estimated SEM for the GVS transformed score $h(y)$, which can be computed using Equations 9 and 10.

Cubic transformation. Moses and Golub-Smith (2011) introduced a scaling method using the cubic function. This method was originally designed to produce scale score distributions with a desired skewness and kurtosis. Moses and Kim (2016) proposed a modified cubic transformation that also can be used to stabilize scale-score CSEMs through numerical optimization (Foi, 2009; Guan, 2009). The mathematical form of the cubic transformed score is

$$sc_{Y,cubic} = \varphi_0 + \varphi_1 y + \varphi_2 y^2 + \varphi_3 y^3. \tag{18}$$

To obtain stabilized scale-score CSEMs, the value of φ_n is numerically solved to minimize the stability function for the scale-score CSEMs. That is, to find the φ_n that minimizes the sum of the squared difference between the cubic transformed score CSEMs at all adjacent scores (Moses & Kim, 2016),

$$\sum_{i=1}^{N}(CSEM_{sc_{Y_i,cubic}} - CSEM_{sc_{Y_{i-1},cubic}})^2, \quad (19)$$

where $CSEM_{sc_{Y,cubic}}$ is the CSEMs of scale scores. As an alternative to using IRT method (Equation 6), when the raw score CSEMs $CSEM_Y$ are known, $CSEM_{sc_{Y,cubic}}$ can be easily estimated by multiplying $CSEM_Y$ by the differentiated cubic transformed score as:

$$CSEM_{sc_{Y,cubic}} = \frac{\partial sc_{Y,cubic}}{\partial Y} CSEM_Y = (\varphi_1 + 2\varphi_2 Y + 3\varphi_3 Y^2) CSEM_Y. \quad (20)$$

Construction of Scale Scores

The target scale was designed to have a mean of 50 and a range of 15 to 85 in integer units. The scale score SEM value was set to 3 for the arcsine and GVS transformations based on Truman L. Kelley's rule of thumb. That is, a 68% confidence interval for scale scores is found by adding ±3 scale score units to the examinee's observed scale score. The standard deviation of the desired scale required by the process of optimization for the cubic transformation was found by selecting standard deviations that led to scale scores with a constant CSEM. When computing the CSEMs for raw scores, 41 quadrature points ranging from -4 to 4 were used. Regarding the GVS and cubic methods, which require raw score CSEMs to generate transformed scales, linear extrapolation methods were used to obtain CSEMs for all possible raw score points (Kolen & Brennan, 2014; Moses and Kim, 2016). Out-of-range scaled scores were truncated; for example, scale scores less than 15 were set equal to 15, and those greater than 85 were set equal to 85. The CSEMs presented in the results section were estimated based on truncated scale scores.

Results

The three scale transformation methods considered in this study were applied to the five pseudo-tests to construct scale scores with relatively stable CSEMs. Table 3 provides summary statistics for scale scores created for the five tests. As can be seen, for each test, the mean for each of the scale scores created by the three transformation methods is close to the target mean score of 50. By comparing the magnitude of the skewness indices, the scale score distributions produced by the cubic approach are closer to being symmetric compared to the scale scores

produced by the arcsine and GVS approaches for Tests 1 and 2. This trend is not apparent for Tests 3 to 5, in which the percentage of points for the MC section dropped to below 50%. All three transformations yield similar reliabilities, with differences less than 0.0001. For each test, the reliabilities of scale scores are also similar to those for the raw scores presented in Table 2.

Figure 2 plots the truncated rounded raw-to-scale score transformations for Tests 1 to 5. For reporting purposes, only the rounded scale scores are shown. Patterns across tests are similar. The arcsine, GVS, and cubic methods all produce curvilinear raw-to-scale score transformation functions that are less steep at the middle scores and relatively steeper at the very high and low scores. The three transformation methods appear to have very similar trends for raw scores ranging from 20 to 60, but the cubic method presents a somewhat different trend than the other two methods at the two extremes of the scale. The arcsine and GVS methods produce a wider and more extreme range of scale scores than the cubic method. Thus, these two methods truncated more points at the lower and upper ends compared to the cubic transformation, which are shown in the figures as points parallel to the horizontal axis.

The CSEMs for the raw and scale scores for the five tests are shown in Figure 3. As expected, the pattern of the CSEMs for the raw scores is bell-shaped. These curves tended to be more symmetric for Tests 3, 4, and 5, in which the percent of points for the MC section are below 50%. As illustrated in the figures, the estimated CSEMs produced by the three transformations are reasonably constant with magnitudes close to 3 for most of the score points near the middle of the score range. However, the resulting scale scores appear to be extreme near the lowest and highest raw scores where the CSEMs for raw scores that are very small. For the scores near the middle of the score range, the GVS and cubic transformations also produce relatively constant CSEMs. The arcsine transformation yields relatively stable CSEMs for Tests 1, 2, and 3. However, for Tests 4 and 5, in which the FR section contributes a larger percent of points, the CSEMs produced by the arcsine transformation fluctuate from 2.8 to 3.2. For very high and very low scores, the CSEMs obtained from the arcsine and GVS transformation drop rapidly; the scale-score CSEMs from the cubic method show a pattern somewhat similar to the raw score CSEMs at the two ends of the scale. These results indicate that the three transformation methods were effective at compressing the CSEMs of most of scores in the middle, but do not sufficiently stretch the scale for extreme scores.

Discussion

Raw scores are typically transformed to scale scores for reporting purposes. Scale transformations are often used to facilitate score interpretation. The current study compared three scale transformation methods designed to simplify score interpretations by stabilizing CSEMs. The transformations considered were the arcsine, GVS, and cubic methods. The applicability of these three scaling procedures was evaluated by applying them to mixed-format tests containing MC and FR items. The effect of mixed-format test composition on the CSEM stabilization was also examined. Five pseudo-tests were constructed based on an operational mixed-format test. These five tests varied in the percentage of points contributed by the MC section. Tests consisting of only dichotomous items or only polytomous items were also included for comparison purposes. IRT was used as the psychometric model to estimate the CSEMs for scale scores.

In general, the results of this study agree with the findings of previous studies (Ban & Lee, 2007; Moses & Kim, 2016; Wang et al., 2000). The raw-to-scale score transformations constructed based on the arcsine, GVS, and cubic procedures were similar at most of the score points in the middle of the score range. The discrepancies among the transformation methods occurred at extreme scores. As expected, the CSEMs for raw scores for all five tests demonstrated bell-shaped curves; that is, the CSEMs for raw scores were larger in the middle and smaller at both ends. These curves tended to be more symmetrical for tests composed of a larger portion of polytomously scored FR items. In summary, the three transformation methods considered in this study successfully stabilized the estimated CSEMs for raw scores in the middle where more examinees were scored. A detailed discussion of each transformation is given next.

The arcsine transformation was effective in stabilizing CSEMs around the middle of the score range. It resulted in more adequate CSEM stabilization for tests consisting of mostly dichotomously scored items. This finding seems reasonable since the arcsine transformation was originally designed for data reflecting binomial assumptions (Kolen & Brennan, 2014).

The scale-score CSEMs produced by the GVS transformation tended to be more stable for most of the score points around the middle of the score range compared to the CSEMs from the arcsine transformation for tests with a larger proportion of polytomous items. The GVS transformation was less sensitive to the test composition compared to the arcsine transformation.

The cubic transformation yielded less extreme scale scores, and less truncation was needed at the two ends than the other two transformations. It was effective in stabilizing CSEMs for most of raw score points near the middle of the score range, and did not appear to be affected by the test composition. However, at the lower and upper ends of the scale, the scale-score CSEMs from the cubic transformation dropped gradually, following the trend of the raw score CSEMs. Moses and Kim (2016) suggested that the cubic transformation has the advantage of symmetry in scale score distributions. However, with the examples used in this study, the symmetry property of the cubic transformation was not significant (Table 3).

In this study, all of the data were based on one test, from which pseudo-tests were formed. The study should be replicated using data from other tests. In addition, the methodology used in this study was based on a unidimensional IRT model. In this study, both the MC and FR items were assumed to assess the same proficiency. Further research should examine the stabilization of CSEMs using a multidimensional IRT model in which the MC items are assumed to assess a different proficiency than the FR items (see Kolen & Lee, 2011).

References

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

Ban, J., & Lee, W. (2007). *Defining a score scale in relation to measurement error for mixed format tests* (CASMA Research Report Number 24). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Cai, L. (2015). *flexMIRT version 3: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Second Edition. Duxbury Press.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York, NY: Macmillan.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics*, *21*, 607-611.

Foi, A. (2009). Optimization of variance-stabilizing transformations. *Preprint*, *2009b*, *94*.

Guan, Y. (2009). Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Statistics and Probability Letters*, *79*, 1621-1629.

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

Jarjoura, D. (1985). Tolerance intervals for true scores. *Journal of Educational Statistics*, *10*, 1–17.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*, 97–110.

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

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice*, *30*(2), 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), 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*(2), 129–140.

Li, D., Woodruff, D., Thompson, T., & Wang, H. (2014, April). *A general method to achieve constant conditional standard error of measurement.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Lord, F. M. (1955). *Estimating test reliability* (Research Bulletin No. RB-55-07). Princeton, NJ: Educational Testing Service.

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

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

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14*, 189–229.

Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (ETS Research Memorandum, ETS RM-11-04). Princeton, NJ: Educational Testing Service.

Moses, T., & Kim, Y. (2016). *Stabilizing conditional standard errors of measurement in scale score transformations*. Manuscript submitted for publication.

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

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*(1), 39–49.

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*(2), 141–162

Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics, 6*, 3–32.

Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement* (ACT Research Report No. 2013-7). Iowa City, IA: American College Testing Program.

Table 1

Number of Items and Possible Points for AP Chemistry and Pseudo-tests

Test	Multiple-Choice Section		Free-Response Section		Total	
	Items	Points	Items	Points	Items	Points
AP Chemistry	75	75	6	61 (10,10,9,15,8,9)	81	136
Test 1	74	74	0	0	74	74
Test 2	60	60	2	20 (10,10)	62	80
Test 3	40	40	4	38 (10,10,9,9)	44	78
Test 4	20	20	6	61 (10,10,9,15,8,9)	26	81
Test 5	0	0	6	61 (10,10,9,15,8,9)	6	61

Table 2

Descriptive Statistics for Raw Scores

Test	Mean	SD	Skewness	Kurtosis	Reliability
Test 1	43.57	13.74	0.03	2.21	0.9318
Test 2	43.09	16.05	0.16	2.12	0.9442
Test 3	37.41	15.81	0.30	2.26	0.9455
Test 4	36.45	17.97	0.29	2.14	0.9529
Test 5	24.71	14.24	0.37	2.18	0.9425

Table 3

Summary Statistics for Scale Scores

Test	Transformation	Mean	SD	Skewness	Kurtosis	Reliability
Test 1	arcsine	50.00	11.42	0.25	2.57	0.9319
	GVS	50.00	11.54	0.28	2.56	0.9325
	cubic	49.99	11.41	0.15	2.30	0.9328
Test 2	arcsine	50.00	12.69	0.31	2.46	0.9433
	GVS	50.01	12.56	0.30	2.49	0.9429
	cubic	49.98	12.43	0.16	2.24	0.9431
Test 3	arcsine	50.00	12.72	0.37	2.60	0.9441
	GVS	50.02	12.80	0.36	2.61	0.9438
	cubic	50.00	12.39	0.29	2.40	0.9444
Test 4	arcsine	50.00	13.60	0.27	2.41	0.9512
	GVS	49.99	13.49	0.25	2.50	0.9505
	cubic	50.01	13.62	0.19	2.30	0.9513
Test 5	arcsine	50.00	12.22	0.26	2.45	0.9401
	GVS	50.00	12.22	0.27	2.56	0.9397
	cubic	49.99	12.76	0.28	2.41	0.9410

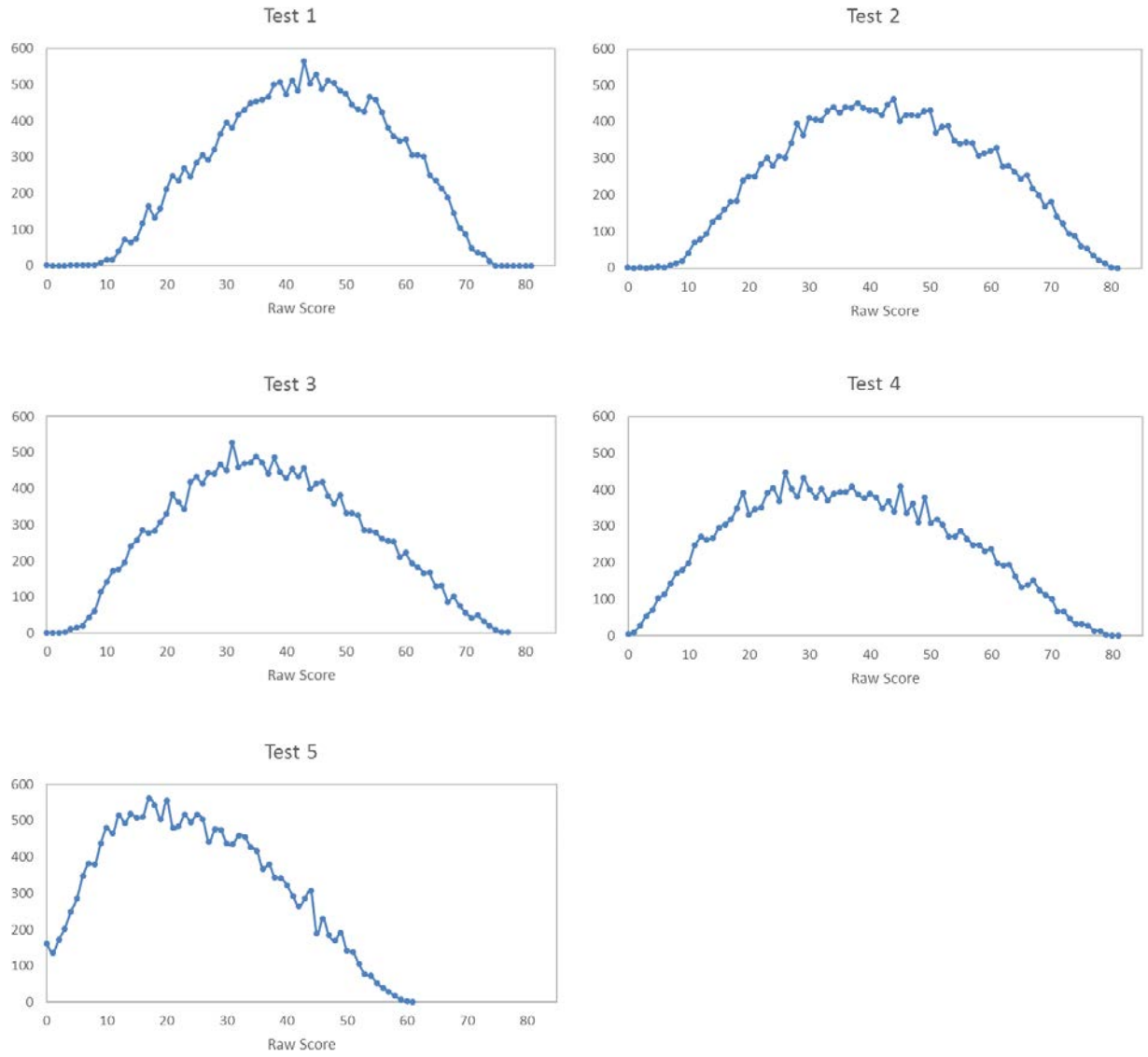

Figure 1. Observed score distributions for Tests 1 to 5.

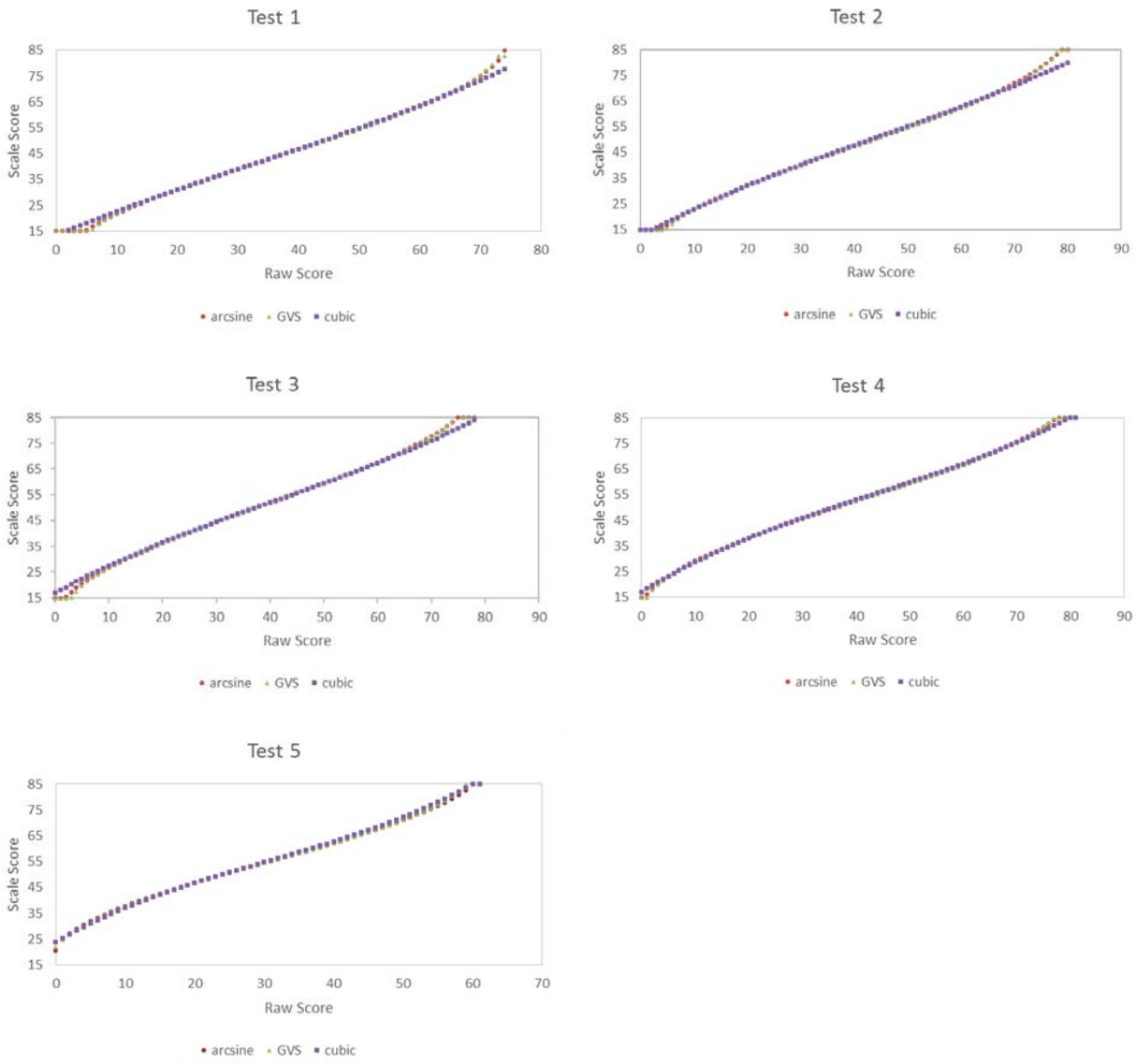

Figure 2. Truncated rounded raw-to-scale score transformations for Tests 1 to 5.

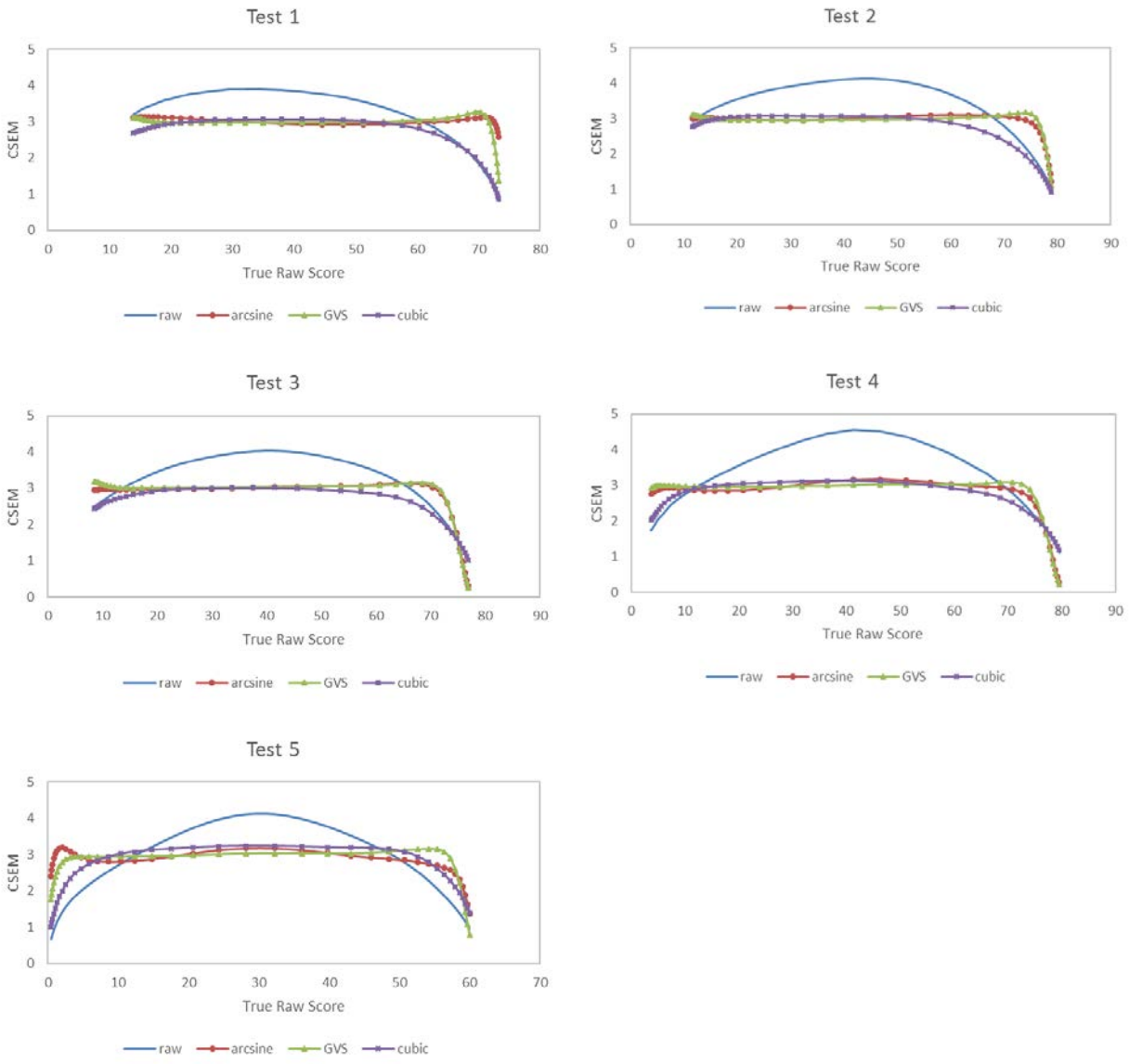

Figure 3. Conditional standard errors of measurement for raw scores and scale scores for Tests 1 to 5.

Chapter 9: A Comparison of IRT Proficiency Interval Estimation Methods for Mixed-Format Tests

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

Abstract

This study reviews and compares various methods for interval estimation of the IRT proficiency parameter for mixed-format tests. The proficiency estimators used to construct intervals include the maximum likelihood (ML) estimator, the method of moments (MM) estimator, the Bayesian expected a posteriori (EAP) estimator with pattern scoring, and the Bayesian EAP estimator with summed scoring. The interval estimation methods investigated are the standard error (SE)-based confidence interval, the ML L_1-based confidence interval, the fiducial interval, the Bayesian credible interval, the percentile interval, and the bias-corrected and acceleration interval. The results of this study suggest that, in general, intervals constructed based on pattern scoring were narrower than the intervals constructed based on summed scoring. Among the methods considered, the L_1-based confidence interval tended to produce the most accurate interval coverage with relatively short length at all proficiency levels.

A Comparison of IRT Proficiency Interval Estimation Methods for Mixed-Format Tests

Introduction

In educational and psychological measurement, examinee proficiency is estimated by administering a set of items to examinees and scoring the performance of the examinees on these items. Under item response theory (IRT), proficiency can be estimated using various methods based on the pattern of examinee item responses or by using a summed score. A point estimator of proficiency, however, only provides partial information by itself; because it does not provide information about how close the point estimator is likely to be to the true value of proficiency. That is, the point estimator alone cannot indicate how much error is involved in the estimation procedure. Interval estimates of examinee proficiency can be used to reflect estimation error by providing a range of values in which lies the specified probability.

There is a large volume of measurement literature on the topic of comparing various proficiency estimators in IRT; however, only limited attention has been paid to deriving proficiency intervals based on these estimators. Lee, Brennan, and Kolen (2006) investigated various procedures for constructing an interval for an examinee's proficiency under the framework of classical test theory; under this theoretical framework, the notion of proficiency is expressed by true score, which is the expected score an examinee would earn over an infinite number of administrations of the test or its parallel forms (Crocker & Algina, 1986). For IRT proficiency estimators, Shyu (2001) compared how different approaches performed in constructing confidence intervals based on two estimators, the maximum likelihood (ML) estimator and the method of moments (MM) estimator, in applications of IRT models. She concluded that the pattern score, maximum-likelihood L_1-based confidence interval was preferable among all confidence intervals studied because it provides more accurate coverage with shorter average length at each proficiency level. It should be noted that Shyu's study is based on tests composed of dichotomous items. It is unknown whether the findings would be the same for mixed-format tests that contain both dichotomous and polytomous items. In addition, Shyu (2001) did not consider interval estimation methods based on widely used estimators, such as the Bayesian expected *a posteriori* (EAP) estimator. These additional estimators should be studied as well.

The primary purpose of this study is to explore different methods in interval estimation for the IRT proficiency parameter using several proficiency estimators with mixed-format tests. In this study, the three-parameter logistic model is used for dichotomously-scored items and the graded response model is used for polytomously scored. The proficiency estimators studied were the ML estimator, the MM estimator, the Bayesian EAP estimator with pattern scoring, and the Bayesian EAP estimator with summed scoring. The interval estimation methods investigated were the standard error (SE)-based confidence interval, the maximum likelihood L_1-based confidence interval, the fiducial interval, the Bayesian credible interval, the percentile interval and the bias-corrected and acceleration interval. The central research question is as follows: How do the interval estimation methods compare in terms of accuracy of coverage and interval width? Simulation techniques were used to address this question.

Method

Data Source

The 2013 forms of the Advanced Placement (AP) English Language and Chemistry mixed-format tests were used as the base for simulation. These two tests differ in section and total test length. There are 54 multiple-choice (MC) items and three free response (FR) items on the English test, with each FR item having a maximum possible score of 9. The maximum summed score for the English test is 81, with the MC section contributing about 67% of the points to the maximum summed score. The Chemistry test is composed of 75 MC items and six FR items. The maximum possible score for the FR items are 10, 10, 9, 15, 8, and 9, respectively. The maximum summed score for the Chemistry test is 136, with the MC section contributing about 55% to the maximum summed score. The AP data were used as the basis for the IRT simulation. It should be noted, however, that IRT is not used to provide scores on the AP, so the results of this study are not directly related to operational AP assessments.

The item parameters of the two tests were calibrated using flexMIRT (Cai, 2015). The MC items were fit with the three-parameter logistic model (Lord, 1980), and the FR items were fit with the graded response model (Samejima, 1997). These estimated item parameters were treated as true item parameters to simulate responses.

Proficiency Estimators

Four proficiency estimators were included in this study: the ML estimator, the MM estimator, the Bayesian EAP estimator with pattern scoring, and the Bayesian EAP estimator

with summed scoring. The MM estimator and Bayesian EAP estimator with summed scoring were obtained using summed scores, which are more easily interpretable by test users; whereas ML estimator and Bayesian EAP estimator with pattern scoring were obtained using examinees' responses to each item, which often leads to greater score precision. In addition, the ML and MM estimators are unbiased estimators; while Bayesian EAP estimators are intended to be biased, but with less overall error.

Maximum likelihood (ML) estimator. Suppose an examinee with ability θ takes a mixed-format test with *n* items, where n_1 of the items are dichotomous and n_2 of the items are polytomous ($n_1 + n_2 = n$). Let u_i be the score for item *i* ($u_i = 0, \ldots m_i$, where m_i designates the maximum possible score for item *i*), and let $U = (u_1, u_2, \ldots, u_n)$ be the vector of responses to all *n* items. If the local independence assumption holds, the likelihood of θ is calculated as:

$$L = L(U|\theta) = \prod_{i=1}^{n_1} P_i^{u_i}(1 - P_i)^{1-u_i} \prod_{j=1}^{n_2} P_{ju_j}, \quad (1)$$

where P_i is the probability of correctly answering a dichotomous item *i*, and P_{ju_j} is the probability of earning a score of u_j for a polytomous item *j*. The value of θ that maximizes the likelihood function is defined as the ML estimator of θ, denoted as $\hat{\theta}_{ML}$. Because lnL is a one-to-one function of *L*, the maximum of *L* and the maximum of lnL occur at the same θ. That is, instead of finding the θ that maximizes *L*, we can find the θ that maximizes lnL. Thus, $\hat{\theta}_{ML}$ can be obtained by solving the following equation, which is set equal to zero:

$$L_1 = \frac{\partial lnL}{\partial \theta} = \sum_{i=1}^{n_1} (u_i - P_i) \frac{P_i'}{P_i(1 - P_i)} + \sum_{j=1}^{n_2} P'_{ju_j}. \quad (2)$$

Method of moments (MM) estimator. The method of moments (Casella & Berger, 1990) estimator is found by treating sample moments as if they were population moments. To apply this to the context of test theory, let an examinee's proportion-correct raw score $\bar{X}$ equal to his or her true proportion-correct score, that is, $\bar{X}$ can be seen as the MM estimator of true proportion-correct score. Because $\bar{X}$ is monotonically related to θ, the MM estimator of θ, denoted as $\hat{\theta}_{MM}$, can be obtained by solving the following equation:

$$\bar{X} = E(\bar{X}|\theta), \quad (3)$$

which can be expressed in the total score metric as equivalent to

$$X = E(X|\theta) = \sum_{i=1}^{n_1} P_i + \sum_{j=1}^{n_2} \sum_{k=0}^{m_j} kP_{jk}, \quad (4)$$

where X is the summed raw score, and P_{jk} is the probability of earning score k for polytomous item j. In practice, $\hat{\theta}_{MM}$ is often used in true score equating (Kolen & Brennan, 2014). It should be noted that for X lower than $\sum_{i=1}^{n_1} c_i$, there is no root for Equation 4. In this case, the MM estimator of θ was set to the lower bound of θ (which is -3.5 in this study).

Bayesian EAP estimator with pattern scoring. Bayesian methods incorporate information of the prior distribution of proficiency to estimate the posterior distribution of proficiency (Bock & Mislevy, 1982). Let $g(\theta)$ be the prior distribution of θ in a population. The joint distribution of examinee response vector U and proficiency θ is the product of likelihood function, shown in Equation 1, and $g(\theta)$:

$$L(U, \theta) = g(\theta)L(U|\theta) = g(\theta)L. \quad (5)$$

The marginal distribution of U, which is the probability of a particular response pattern in the population, is found by integrating $L(U, \theta)$ over θ:

$$L(U) = \int_{-\infty}^{\infty} g(\theta)Ld\theta. \quad (6)$$

Based on Bayes theorem, the posterior distribution of θ given examinee response vector U can be expressed as

$$\Pr(\theta|U) = \frac{L(U, \theta)}{L(U)} = \frac{g(\theta)L}{\int_{-\infty}^{\infty} g(\theta)Ld\theta}. \quad (7)$$

The Bayesian EAP estimator with pattern scoring, denoted as $\hat{\theta}_{EAP}$, is defined as the mean of the posterior distribution of θ. It is computed by multiplying the preceding equation by θ and integrating over θ. In practice, θ is often approximated with discrete quadrature points and $g(\theta)$ is approximated with quadrature weights, and thus, $\hat{\theta}_{EAP}$ is found by

$$\hat{\theta}_{EAP} = \frac{\sum_{\theta} \theta g(\theta)L}{\sum_{\theta} g(\theta)L}. \quad (8)$$

Forty quadrature points were used in this study.

Bayesian expected a posteriori estimator with summed scoring (SEAP). Thissen and Orlando (2001, p. 121) suggested a Bayesian EAP estimator based on summed scoring, denoted as $\hat{\theta}_{SEAP}$,

$$\hat{\theta}_{SEAP} = \frac{\sum_{\theta} \theta \Pr(X|\theta) L}{\sum_{\theta} \Pr(X|\theta) L}, \tag{9}$$

where $\Pr(X|\theta)$ is the conditional probability distribution of raw scores. For tests containing both dichotomous and polytomous items, $\Pr(X|\theta)$ can be obtained using an extension of the Lord-Wingersky algorithm provided by Hanson (1994), Thissen, Pommerich, Billeaud, and Williams (1995), and Wang, Kolen, and Harris (2000).

Interval Estimation

Several methods for constructing an interval with a nominal confidence level of $100(1-\alpha)\%$ for θ were considered in this study. Some of these methods , such as the SE-based confidence interval, the L_1-based confidence interval, the fiducial interval, and the credible interval, compute intervals using analytic methods by making assumptions about the forms of underlying population. Other intervals, including the percentile interval and the bias-corrected and accelerate interval, were found using an empirical approach, which involves resampling from the given response data. For different proficiency estimators, different methods were used: the SE-based and L_1-based confidence intervals were computed for $\hat{\theta}_{ML}$; the SE-based confidence intervals and fiducial interval were computed for $\hat{\theta}_{MM}$; and the Bayesian credible interval was constructed for $\hat{\theta}_{EAP}$ and $\hat{\theta}_{SEAP}$. The percentile interval and bias-corrected and accelerate interval were applied to all of the proficiency estimators.

SE-based confidence interval. An SE-based confidence interval is constructed by assuming that errors of measurement are normally distributed (Feldt & Brennan, 1989). The approximate $100(1-\alpha)\%$ confidence intervals for θ is

$$\left(\hat{\theta} + z_{\alpha/2} SE(\hat{\theta}),\ \hat{\theta} - z_{\alpha/2} SE(\hat{\theta})\right), \tag{10}$$

where $z_{\alpha/2}$ is the $100(\alpha/2)$th percentile point of the standard normal distribution, and $SE(\hat{\theta})$ is the standard error (SE) of $\hat{\theta}$. An asymptotic expression for SE of $\hat{\theta}_{ML}$, is the square root of the variance of $\hat{\theta}_{ML}$ and can be expressed as

$$SE(\hat{\theta}_{ML}) = \sqrt{var_{ML}} = \sqrt{\frac{1}{I(\theta)}}, \tag{11}$$

where $I(\theta)$ is the information of the mixed-format test, computed as

$$I(\theta) = \sum_{i=1}^{n_1} \frac{(P_i')^2}{P_i(1-P_i)} + \sum_{j=1}^{n_2} \sum_{k=0}^{m_j} \frac{(P_{jk}')^2}{P_{jk}}. \quad (12)$$

whereP_i' and P_{jk}' denotes the first derivative of P_i and P_{jk}, respectively. For $\hat{\theta}_{MM}$, the SE($\hat{\theta}_{MM}$) is

$$SE(\hat{\theta}_{MM}) = \sqrt{\sum_{i=0}^{N} \theta \Pr(X=j|\theta) - \left[\sum_{i=0}^{N} \theta \Pr(X=j|\theta)\right]^2}, \quad (13)$$

where N is the maximum possible score for the mixed-format test.

L_1-based confidence interval. Lloyd (1984) stated that the first derivative of the log-likelihood function (L_1 in Equation 2) is approximately normally distributed with a mean of zero and a variance equal to the negative expectation of the second derivative of the log-likelihood function. When applied to the context of IRT proficiency estimation, as done by Shyu (2001), L_1 can be seen as a random variable of n independent random variables u_i $(i = 1, \ldots, n)$ with θ fixed. Thus, L_1 is approximately normal with a mean of zero and a variance equal to the test information function $I(\theta)$ (Kendall & Stuart, 1977). Using this normal approximation, a realized $100(1-\alpha)\%$ confidence interval for L_1 can be found as $(z_{\alpha/2}\sqrt{I(\theta)}, -z_{\alpha/2}\sqrt{I(\theta)})$. Since L_1 is monotonically related to θ, the two endpoints of the L_1-based confidence interval of θ can be obtained by solving the following equation,

$$L_1 \pm z_{\alpha/2}\sqrt{I(\theta)} = 0. \quad (14)$$

Fiducial interval. A fiducial interval is produced based on the conditional probability distribution of raw scores $\Pr(X|\theta)$. For an observed summed score x_0 from the given response pattern, the two endpoints of a $100(1-\alpha)\%$ confidence interval of θ can be found by solving the following two equations

$$\Pr(X \leq x_0|\theta) = \sum_{i=0}^{x_0} \Pr(X=i|\theta) = \alpha/2 \quad (15)$$

and

$$\Pr(X \geq x_0|\theta) = \sum_{i=x_0}^{N} \Pr(X=i|\theta) = \alpha/2. \quad (16)$$

Credible interval. A credible interval is obtained based on Bayesian inference. A $100(1-\alpha)\%$ credible interval is the interval from the $100(\alpha/2)$th to $100(1-\alpha/2)$th percentile of the posterior distribution of θ. For $\hat{\theta}_{EAP}$, the posterior distribution of θ is provided in Equation 7. For $\hat{\theta}_{SEAP}$, the posterior distribution of θ can be found by replacing $g(\theta)$ in Equation 7 with $\Pr(X|\theta)$.

Percentile interval. The percentile interval is typically generated based on bootstrap procedures by resampling with replacement from the given data. In this study, it is obtained through simulation techniques by simulating a distribution of $\hat{\theta}$. With the known item parameters and an estimate $\hat{\theta}$ based on the given response pattern, the percentile interval is found by the following four steps:

1. Compute cumulative probabilities of scoring at or below category k for each item i based on the obtained $\hat{\theta}$, $P_{ik}^{*}(\hat{\theta})$, where $k = 0, \ldots, m_i$ and $i = 1, \ldots, n$.
2. Obtain a response vector, $U = (u_1^{*}, \ldots u_n^{*})$, where u_i^{*} is a simulated response for item i generated by comparing a random number from a uniform distribution to $P_{ik}^{*}(\hat{\theta}), k = 0, \ldots, m_i$. If the random number is less than $P_{i(k+1)}^{*}(\hat{\theta})$ and larger than or equal to $P_{ik}^{*}(\hat{\theta})$, the assigned score to item i is k.
3. Compute $\hat{\theta}^{*}$ based on the response vector obtained in step 2 using the chosen proficiency estimation method for estimation of $\hat{\theta}$.
4. Repeat steps 2 to 3 1,000 times to obtain a distribution of $\hat{\theta}$.

According to Efron and Tibshirani (1993), a $100(1-\alpha)\%$ percentile interval for θ is given by $(\hat{\theta}_{\alpha/2}^{*}, \hat{\theta}_{1-\alpha/2}^{*})$, and the two endpoints $\hat{\theta}_{\alpha/2}^{*}$ and $\hat{\theta}_{1-\alpha/2}^{*}$ are the $100(\alpha/2)$th and the $100(1-\alpha/2)$th percentiles of the simulated distribution of $\hat{\theta}$, respectively.

Bias-corrected and accelerated (BCa) interval. The BCa interval (Efron & Tibshirani, 1993) requires more extensive computation than the percentile interval. A $100(1-\alpha)\%$ BCa interval for θ is given by $(\hat{\theta}_{\alpha_1}^{*}, \hat{\theta}_{2}^{*})$, where $\hat{\theta}_{\alpha_1}^{*}$ and $\hat{\theta}_{\alpha_2}^{*}$ are the $100(\alpha_1)$th and $100(\alpha_2)$th percentiles of the simulated distribution of $\hat{\theta}$, respectively. α_1 and α_2 are computed as

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{\alpha}(\hat{z}_0 + z_{\alpha/2})}\right) \tag{17}$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{\alpha}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right), \quad (18)$$

where Φ is the standard normal cumulative distribution function. In Equations 17 and 18, the bias-corrected parameter $\hat{z}_0$ is the proportion of replications less than the original estimate ($\hat{\theta}$), and the acceleration parameter $\hat{\alpha}$ measures how quickly the standard error is changing on the normalized scale. They are computed as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_r^* < \hat{\theta}\}}{1000}\right) \quad (19)$$

and

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6[\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2]^{3/2}}, \quad (20)$$

where r is from 1 to 1000, n is the total number of items, $\hat{\theta}_{(i)}$ is the estimate of θ obtained using the original sample with the ith item omitted, and $\hat{\theta}_{(.)}$ is the mean of $\hat{\theta}_{(i)}$ ($\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$).

True Proficiency Levels

Eleven proficiency levels ranging from -2.5 to 2.5 with an increment of 0.5 were examined in the current study. The upper and lower bounds for θ estimates were set as -3.5 and 3.5, respectively. The lower limits of the constructed interval that were less than -3.5 were set to -3.5, and upper limits that were greater than 3.5 were set to 3.5.

Numerical Procedure

An approximation method must be used to obtain $\hat{\theta}_{ML}$, $\hat{\theta}_{MM}$, the endpoints of the L_1-based confidence interval, and the endpoints of the fiducial interval. The most popular approximation method is Newton-Raphson. However, the final solution with this method is highly dependent on a fine-tuned initial value. When a poor initial value is used, it may cause divergence and find a local maximum rather than a universal maximum. A combination of the Bisection and Newton-Raphson algorithm (Press, Teukolsky, Vetterling, & Flannery, 1996) that can handle cases of divergence efficiently was used in this study to find $\hat{\theta}_{ML}$ and $\hat{\theta}_{MM}$. In addition, the Newton-Raphson method requires calculating the first derivative of the evaluated function, but calculation of the first derivative for the equations used to find the endpoints of the L_1-based confidence interval (Equation 4) and fiducial interval (Equations 5 and 6) is very complicated; to accommodate, an exhaustive search algorithm was adopted. The concept of

exhaustive search (Wothke, Burket, Chen, Gao, Shu, & Chia, 2011) is straightforward, which is to try every possible value within a certain range and select the one with the most satisfying result. As long as enough precision is employed, the exhaustive search algorithm is guaranteed to find the target value. In this study, 7,001 θ points from -3.5 to 3.5 with an increment of 0.001 were used to search for an accurate estimate. One drawback of this algorithm is that it is computationally-intensive and relatively slow.

Simulation Procedures

For each test included in this study, the following simulation steps were used:

1. Compute cumulative probabilities of scoring at or below category k for each item i based on a true θ point, $P_{ik}^{*}(\theta)$, where $k = 0, \dots, m_i$ and $i = 1, \dots, n$.
2. Obtain a response vector, $U = (u_1, \dots u_n)$, where u_i is a simulated response for item i generated by comparing a random number from a uniform distribution to $P_{ik}^{*}(\theta), k = 0, \dots, m_i$. If the random number is less than $P_{i(k+1)}^{*}(\theta)$ and larger than or equal to $P_{ik}^{*}(\theta)$, the assigned score to item i is k.
3. Based on the response vector obtained in step 2, compute $\hat{\theta}_{ML}$, $\hat{\theta}_{MM}$, $\hat{\theta}_{EAP}$, and $\hat{\theta}_{SEAP}$ and their corresponding intervals using various methods.
4. Repeat steps 2 to 3 1,000 times.
5. Repeat steps 1 to 4 to all eleven true θ points.

Evaluation Criteria

Intervals constructed using the methods described in the previous section were computed at a level of 95% ($\alpha = 0.05$). To find the interval estimation method that yields the most accurate interval, coverage, which is the proportion of counts out of the total replications in which the interval contains the true proficiency, at each proficiency level was compared. The rule of standard error of the empirical size (SEES) suggested by Shyu (2001) was adopted to examine the performance of coverage at each proficiency level. The SEES for a 95% interval was computed as $\sqrt{(0.95 * 0.05)/1000}$, which equals 0.68%. The coverage or empirical size of the constructed interval within two SEES was considered as adequate. That is, the coverage between 93.64 and 96.36 is considered adequate. In addition, the average length of each interval at each proficiency level was also compared. An interval estimation method that produced relatively short interval in length with the coverage probability close to the nominal level of 95% was preferred.

Results

The differences between the nominal size of 95% and the empirical sizes of intervals constructed using the ML estimator, the MM estimator, the Bayesian EAP estimator with pattern scoring, and the Bayesian EAP estimator with summed scoring are presented in Tables 1 to 4, respectively. The differences are bolded if they are within two SEES. Figures 1 to 6 show the average lengths of the 95% interval constructed based on all four estimators. The results for intervals were first examined separately under each estimator, and then compared.

Under ML estimation, Table 1, the L_1-based confidence interval performs the best in terms of coverage at most of the proficiency levels for both the English and Chemistry tests, followed by the SE-based confidence interval. The percentile interval and BCa interval performed poorly in both tests, especially at very low proficiency levels. As shown in Figure 1, the pattern of average lengths of the intervals produced by the four methods are similar and displays a U-shaped pattern with lower average length in the middle and higher length in the two ends. BCa interval has slightly shorter average length across all proficiency levels compared to other three intervals.

In general, the empirical sizes for confidence intervals constructed using the MM estimator shown in Table 2 were off more from the nominal level than the intervals constructed using the ML estimator. Among the four intervals, in terms of coverage, the fiducial interval performed better and varied less across proficiency levels than the other three intervals. The SE-based confidence interval and BCa interval had poorer coverage. For average lengths of constructed intervals, the patterns of intervals obtained from the MM estimator were similar to those obtained from the ML estimator. In Figure 2, it is obvious that the fiducial interval yields larger average length compared to other intervals form MM estimation, which indicates that its good coverage performance may be the result of longer intervals. The trends of average length of the SE-based confidence interval, the percentile interval, and the BCa interval under the MM estimation were consistent with those under the ML estimation.

Using the Bayesian EAP estimator with pattern scoring (Table 3) and summed scoring (Table 4), credible intervals worked well in covering the true proficiency for levels at the middle proficiency levels, especially for the Chemistry test as shown in Table 3. In terms of coverage, the percentile and BCa intervals based on Bayesian EAP estimator showed the poorest performance compared to those based on the ML or MM estimators at nearly all levels. Because

pattern scoring incorporated more information into the estimation process than summed scoring, the empirical sizes of the intervals shown in Figures 3 and 4 computed from Bayesian EAP estimator with pattern scoring were closer to the nominal level compared with the ones from summed scoring. In terms of average length, the patterns for the intervals under Bayesian EAP estimation were similar to the ones under ML and MM estimation. The difference at each level across methods were smaller than 0.2.

Comparing the intervals constructed under four estimation methods altogether, it is clear that estimators using pattern soring such as the ML estimator and Bayesian EAP estimator with pattern scoring yielded better coverage performance than estimators using summed scoring. The empirical sizes of the intervals computed using analytic methods were closer to the nominal size than the ones computed using the empirical methods. Among the analytic methods considered in this study, the L_1-based confidence interval using the ML estimator yielded the best results in terms of both coverage and length at all levels, followed by the credible interval using Bayesian EAP estimator with pattern scoring. The performance of the SE-based confidence interval was not consistent across levels. Figures 5 and 6 summarize the average lengths of intervals for θ using analytic and empirical methods, respectively. At nearly all levels, the average interval lengths for each method were smaller for the Chemistry test, which is longer in length. Among all the analytic methods considered, the fiducial and SE-based intervals under the MM estimation had larger average length across levels. The credible interval under the Bayesian EAP estimator with patterned scoring produced the shortest average length.

Discussion

The primary goal of any test is to accurately measure an examinee's ability based on his or her performance on that test. Therefore, it is important to identify a method that provides better interpretation of an examinee's proficiency. This study focused on investigating various methods in interval estimation for the proficiency parameter for mixed-format tests under the framework of IRT. The four estimators used to construct intervals for proficiency included the ML estimator, the MM estimator, the Bayesian EAP estimator with pattern scoring, and the Bayesian EAP estimator with summed scoring. Analytic and empirical methods, including the L_1-based confidence interval, the fiducial interval, the Bayesian credible interval, the percentile interval, and the bias-corrected and acceleration interval, were considered for constructing

intervals for proficiency. Data simulated from real tests were used to compare the performance of these intervals at eleven proficiency levels.

In general, the results of this study were consistent with previous studies based on tests with only dichotomous items (Shyu, 2001). This study also showed that confidence intervals constructed using the ML estimator performed better in terms of coverage than the ones constructed based on the MM estimator. Using the ML estimator, the L_1-based confidence interval outperformed the SE-based confidence interval at nearly all levels. The fiducial interval based on the MM estimator produced good coverage but with larger average length. Overall, the analytic methods yielded better coverage performance than the empirical methods. In addition, this study broadened the findings for interval estimation for IRT proficiency by considering Bayesian credible intervals and empirical intervals computed using Bayesian EAP estimator with both pattern scoring and summed scoring. Based on the data used in this simulation study, the present study indicated that intervals constructed based on pattern scoring provided more precise results than the ones constructed based on summed scoring. The credible interval produced results with more accurate coverage with shorter length for proficiency levels around 0.

The present study was based on data simulated from two particular mixed-format tests. The results should be generalized only to mixed-format tests with similar statistical item characters and composition with the tests used in this study. In addition, since this study only considered the intervals constructed based on four proficiency estimators, other proficiency estimation methods such as test characteristic function with summed scoring, Yen's (1984) maximum likelihood estimation, and Bayesian maximum *a posteriori* estimator should be researched as well.

References

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

Cai, L. (2015). *flexMIRT version 3: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Casella, G., & Berger, R. L. (1990). *Statistical Inference*, Belmont, CA: Duxbury Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modem test theory.* New York: Holt, Rinehart, and Winston.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistician Association, 82,* 171-200.

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

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). New York: Macmillan.

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

Lee, W., Brennan, R. L., & Kolen, M. J. (2006). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics, 31*(3), 261-281.

Lloyd, E. (1984). Maximum likelihood estimates. In W. Ledermann (Ed.), *Handbook of applicable mathematics* (pp. 283-354). New York: Wiley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*(4), 453-461.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1996). *Numerical recipes in C* (Vol. 2). Cambridge: Cambridge university press.

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

Shyu, C. Y. (2001). *Estimating error indexes in estimating proficiencies and constructing confidence intervals in item response theory.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3009638).

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Ed.), *Test scoring* (pp. 73-140). Mahwah, NJ: Erlbaum.

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*(1), 39–49.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93-111.

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*(2), 141–162

Wothke, W., Burket, G., Chen, L. S., Gao, F., Shu, L., & Chia, M. (2011). Multimodal Likelihoods in Educational Assessment Will the Real Maximum Likelihood Score Please Stand Up? *Journal of Educational and Behavioral Statistics, 36*(6), 736-754.

Table 1

Differences of Nominal Size 95% and Empirical Sizes of the Intervals Constructed Using the ML Estimator

θ Level	English				Chemistry			
	L_1	SE	Percentile	BCa	L_1	SE	Percentile	BCa
-2.5	**-0.2**	-1.4	-1.4	**-1.3**	**0.9**	2.3	3.0	2.6
-2	**-0.4**	**-0.6**	-1.8	-1.5	**-0.3**	**0.3**	-2.1	1.9
-1.5	**-0.9**	-1.9	-2.7	-1.8	**-0.2**	**0.1**	**-0.4**	1.9
-1	**-0.7**	**-1.2**	-2.2	**1.2**	**0.0**	**-0.2**	**-0.5**	2.3
-0.5	**-0.3**	**-0.1**	-1.4	**0.8**	**-0.2**	**0.9**	**0.6**	**0.8**
0	**0.4**	**0.7**	**-0.9**	**-0.1**	**-0.3**	**0.2**	**-0.6**	**0.5**
0.5	**0.1**	**0.3**	**-0.3**	**0.2**	**-0.9**	-1.4	-1.4	**0.9**
1	**0.8**	**1.1**	**-0.9**	**-0.7**	**-0.7**	**-0.4**	**-0.9**	**0.6**
1.5	**0.2**	**0.1**	-2.0	1.9	**-0.1**	**-0.2**	**-0.6**	**-1.0**
2	**0.6**	**0.7**	**0.1**	**0.4**	**-0.8**	**-0.8**	**-0.5**	**0.4**
2.5	**-0.1**	**-0.5**	-1.5	**-1.0**	-2.0	**-0.5**	-2.8	**0.7**

Table 2

Differences of Nominal Size 95% and Empirical Sizes of the Intervals Constructed Using the MM Estimator

θ Level	English				Chemistry			
	SE	Fiducial	Percentile	BCa	SE	Fiducial	Percentile	BCa
-2.5	-5.5	**0.3**	-3.5	3.0	-9.0	1.7	2.2	-2.4
-2	**-1.0**	2.2	**-0.9**	**-1.2**	**-1.0**	**0.7**	**0.2**	1.8
-1.5	-3.9	**-0.7**	-2.0	2.6	**-0.2**	**0.7**	**-0.4**	2.6
-1	**-1.3**	**0.7**	**0.7**	**1.3**	**-0.3**	1.5	**-0.5**	2.8
-0.5	**-0.5**	**-1.3**	**0.5**	**0.2**	-1.8	**1.0**	**0.7**	1.7
0	1.9	**1.2**	2.6	1.8	**-0.1**	**-0.1**	**-0.3**	**1.1**
0.5	-2.0	**-0.9**	**-1.2**	-1.5	-1.9	-1.9	-2.1	**1.0**
1	**1.2**	**1.2**	1.5	1.8	**0.1**	**0.1**	**-0.8**	**0.9**
1.5	**-0.4**	**0.4**	**0.4**	1.7	**-0.7**	**0.1**	**-0.9**	1.6
2	-5.7	**0.6**	**-1.1**	-2.0	1.7	2.1	**-0.2**	**0.9**
2.5	-7.8	**0.2**	**-1.3**	-2.7	-4.6	**0.0**	-2.8	**0.6**

Table 3

Differences of Nominal Size 95% and Empirical Sizes of the Intervals Constructed Using the Bayesian EAP Estimator with Pattern Scoring

θ Level	English			Chemistry		
	Credible	Percentile	BCa	Credible	Percentile	BCa
-2.5	-6.5	-8.6	-5.0	**0.0**	-4.1	-7.3
-2	-1.5	-6.1	-4.1	**-1.3**	-3.9	-5.2
-1.5	-1.6	-5.7	3.0	**-0.3**	-1.4	-2.2
-1	**0.3**	**-0.1**	-2.3	**0.4**	**0.1**	**1.0**
-0.5	**-0.5**	**-0.8**	**-1.1**	**0.8**	**0.7**	1.4
0	**0.4**	**-0.6**	**0.2**	**0.5**	**0.0**	1.8
0.5	**0.5**	**0.6**	**-0.8**	**-1.0**	**-0.6**	2.2
1	**1.2**	**0.4**	**1.2**	**-0.2**	**0.5**	3.7
1.5	**0.7**	-2.1	2.5	**-0.8**	-1.8	3.9
2	-1.8	-11.1	3.0	**-1.0**	-3.2	3.9
2.5	-6.1	-15.0	-3.2	-3.0	-8.2	3.4

Table 4

Differences of Nominal Size 95% and Empirical Sizes of the Intervals Constructed Using Bayesian EAP Estimator with Summed Scoring

θ Level	English			Chemistry		
	Credible	Percentile	BCa	Credible	Percentile	BCa
-2.5	-4.9	-16.5	-3.2	-1.8	**-1.1**	-2.0
-2	-3.3	**-1.1**	3.6	**-0.5**	1.7	1.9
-1.5	**0.1**	**0.5**	-3.1	-1.6	**0.4**	3.3
-1	**-0.4**	**0.3**	2.6	1.5	**0.6**	3.9
-0.5	**-0.4**	-5.2	3.0	**0.3**	**-0.1**	4.0
0	**0.2**	-3.6	-3.2	**0.2**	**-0.8**	3.7
0.5	**-1.2**	-7.6	2.8	**-0.8**	-3.9	4.0
1	**0.2**	-9.1	3.5	**-0.3**	-7.3	4.0
1.5	**-0.6**	-10.0	1.8	**-0.9**	-14.0	3.2
2	**-1.3**	-32.0	3.1	**-1.1**	-24.5	2.3
2.5	-5.8	-52.2	3.6	-4.6	-34.7	3.8

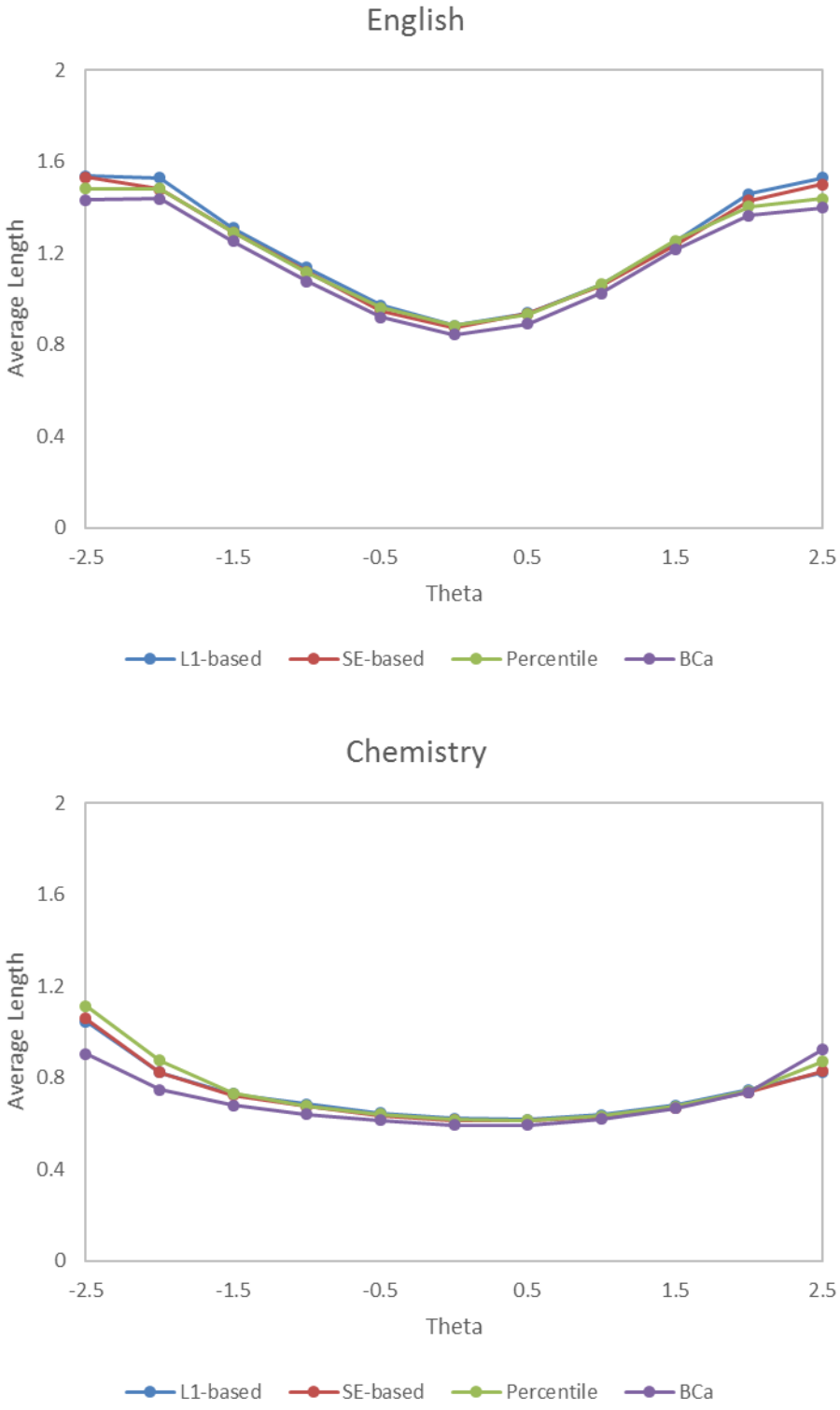

Figure 1. Average length of 95% intervals for θ using ML estimator.

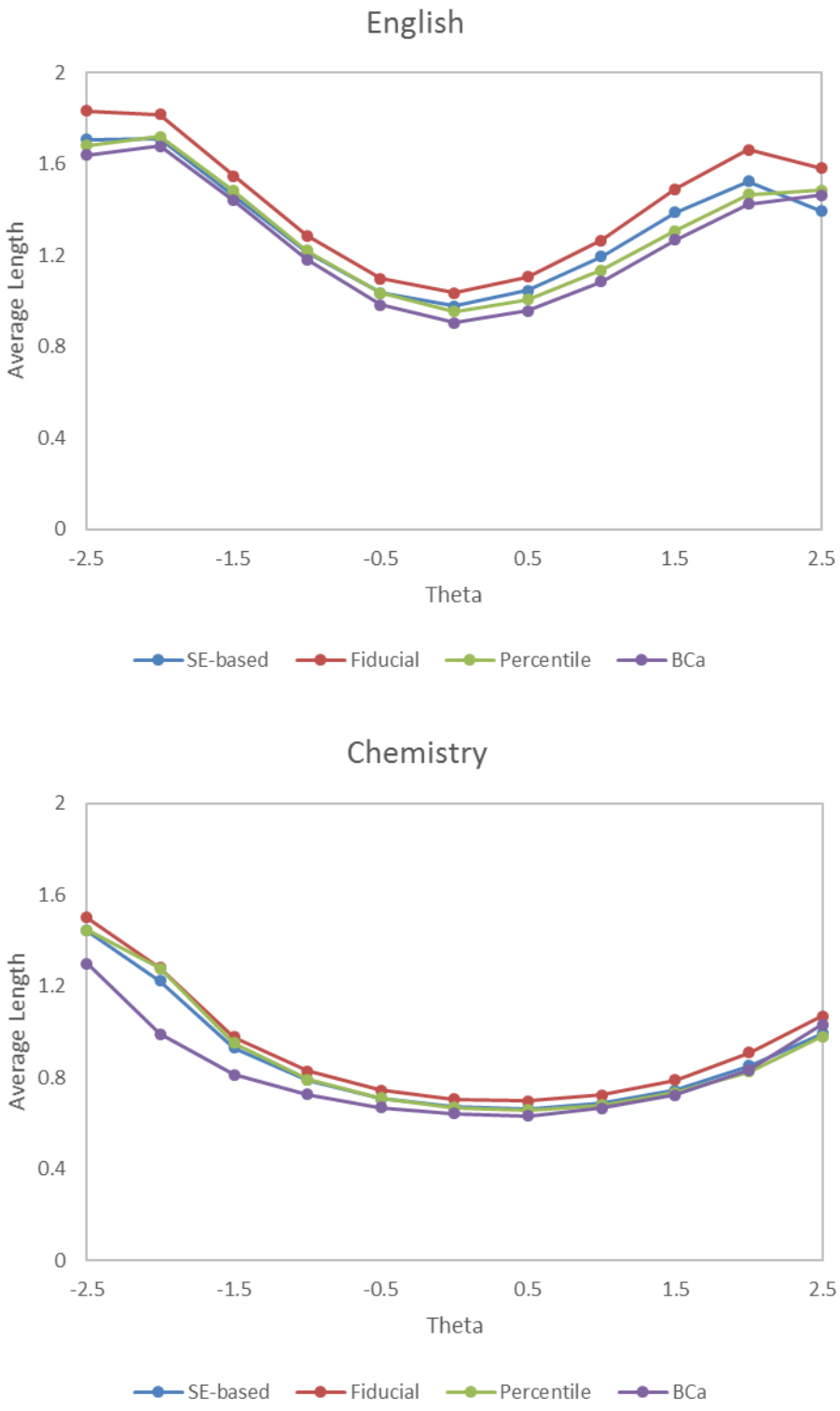

Figure 2. Average length of 95% intervals for θ using MM estimator.

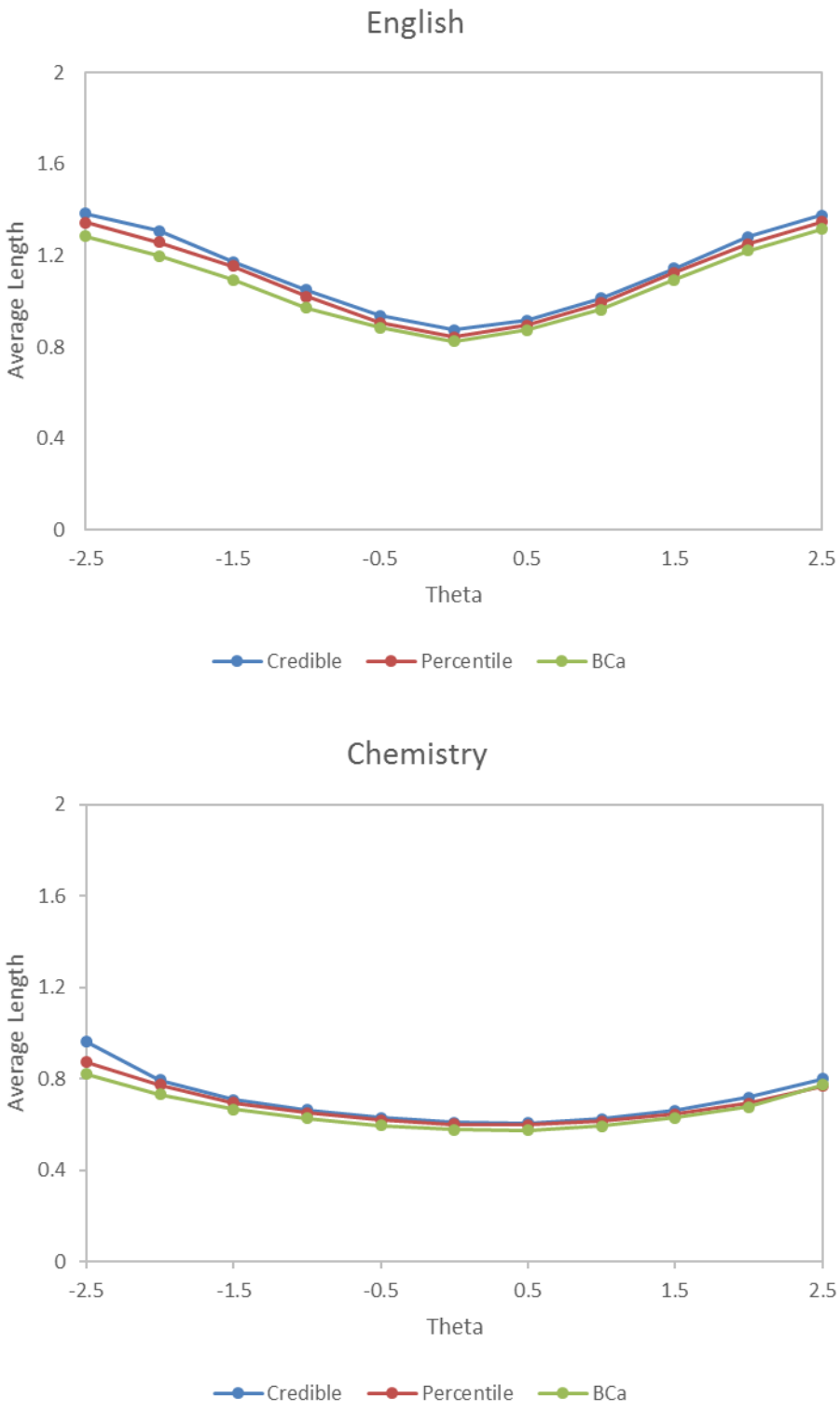

Figure 3. Average length of 95% intervals for θ using Bayesian EAP estimator with pattern scoring.

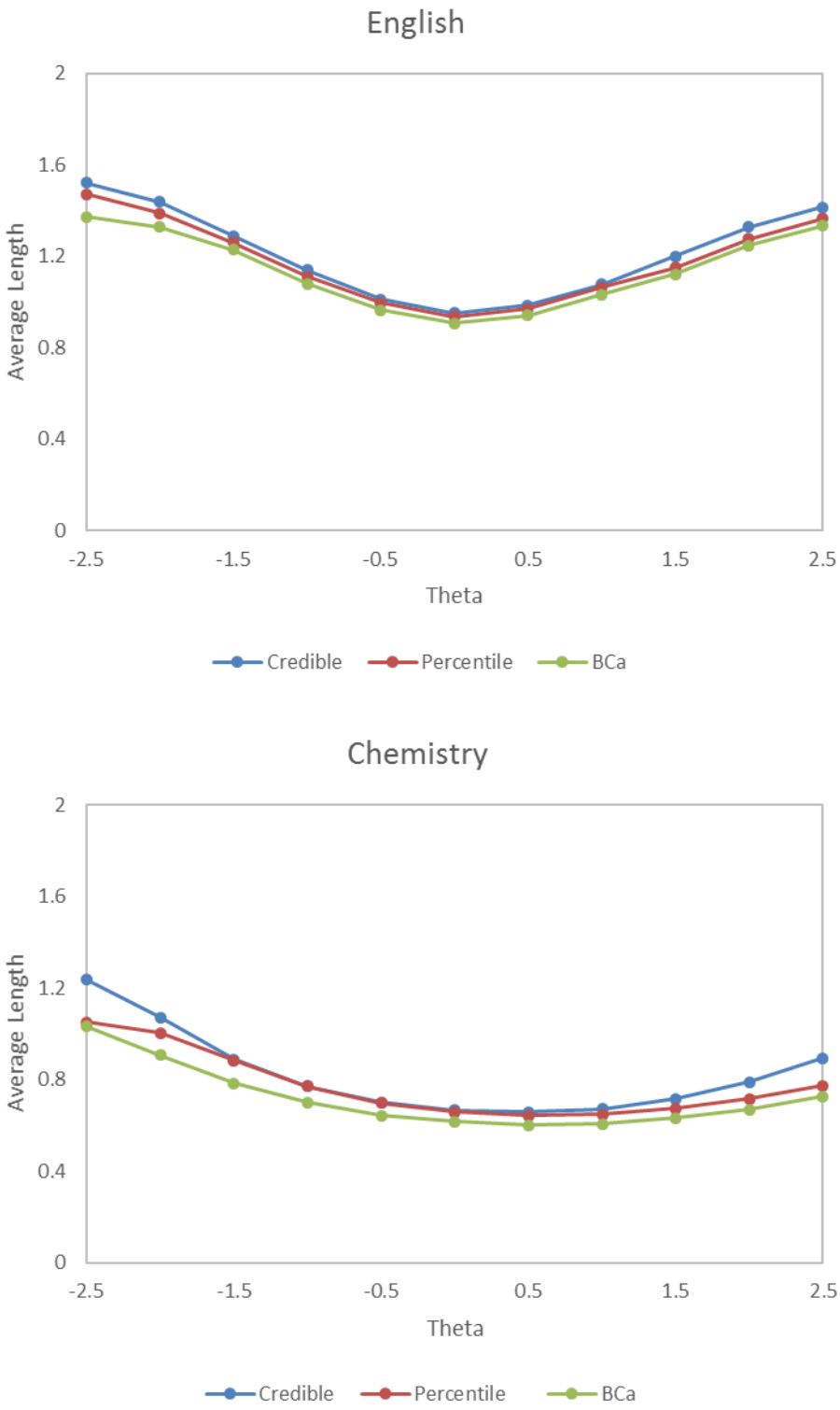

Figure 4. Average length of 95% intervals for θ using Bayesian EAP estimator with summed scoring.

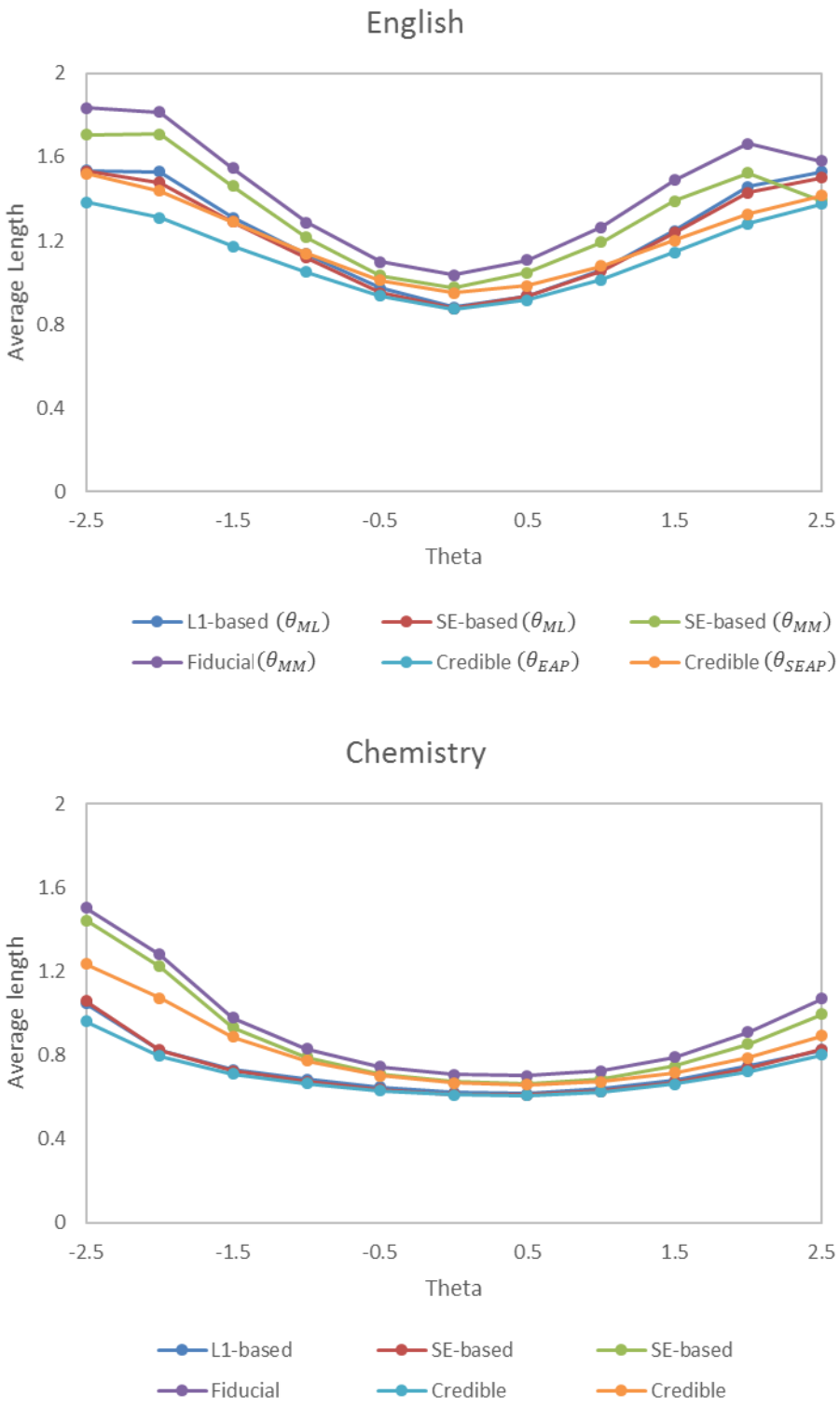

Figure 5. Average length of 95% intervals for θ using analytic methods.

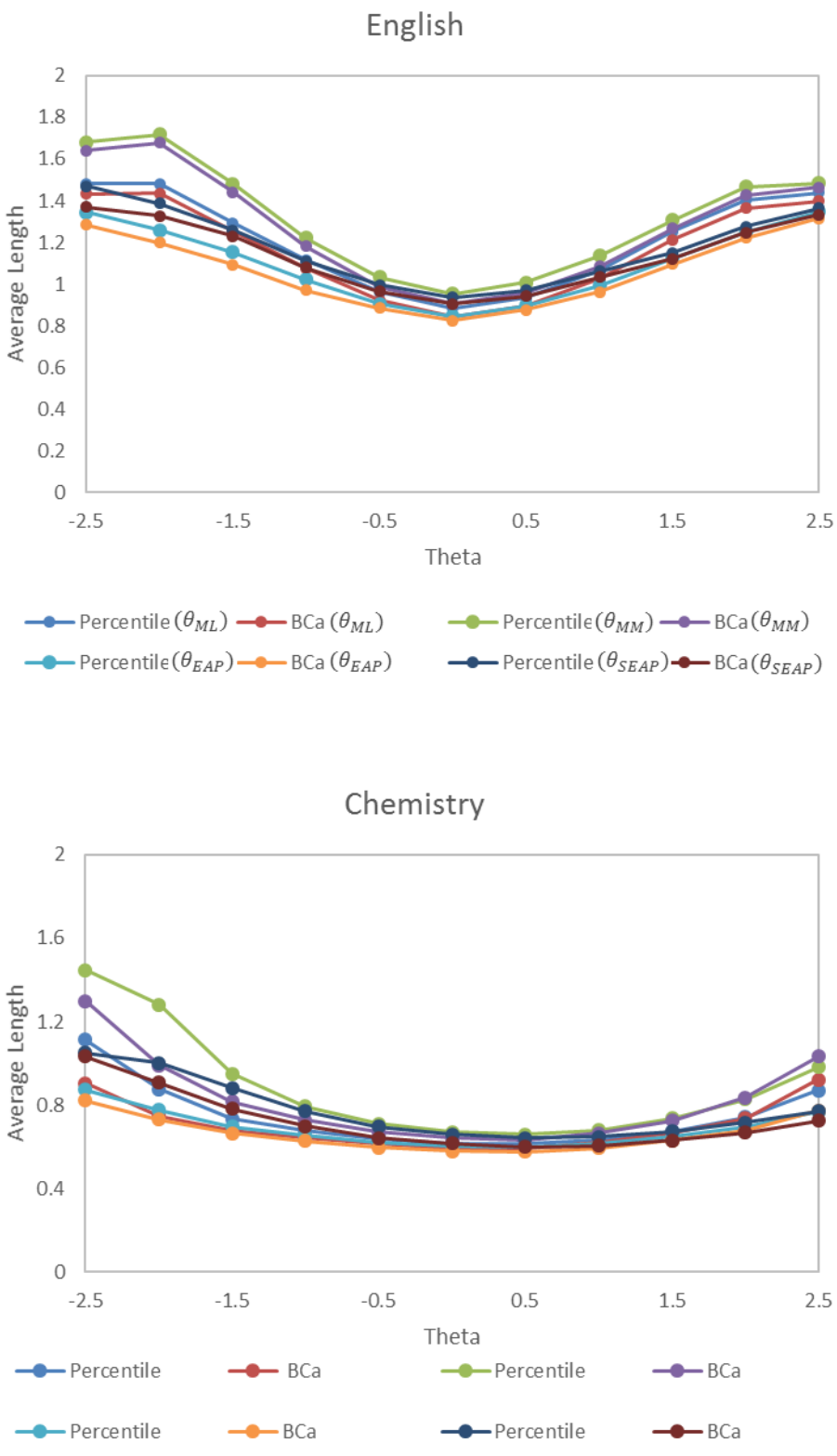

Figure 6. Average length of 95% intervals for θ using empirical methods.