Center for Advanced Studies in
Measurement and Assessment

CASMA Monograph

Number 2.2

Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 2)

Michael J. Kolen
Won-Chan Lee
(Editors)

December, 2012

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 2

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

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

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

Volume 2 contains 6 chapters. Chapter 1 provides an overview and describes methodology that is common across many of the chapters. In addition, it highlights some of the methodological issues encountered and some of the major findings.

Chapter 2 is a simulation study that investigates the feasibility of equating mixed-format test forms using a common-item set that consists solely of multiple-choice (MC) items. Chapter 3 considers the effect of group differences on equating results for mixed-format test forms based on two different designs and criteria. Chapter 4 examines the usefulness of matched samples equating. Chapter 5 introduces and investigates an equating method based on the use of a simple structure multidimensional item response theory (IRT) model. Chapter 6 investigates IRT item fit statistics.

We thank Bradley Brossman (now at the American Board of Internal Medicine), Sarah Hagge (now at the National Council on State Boards of Nursing), Yi He (now at ACT Inc.), Sonya Powers (now at Pearson), and Wei Wang (a University of Iowa graduate student currently supported by an Educational Testing Service Gulliksen Fellowship) for their work. We also thank graduate student Jaime Peterson for her effort in producing this volume.

We thank Robert L. Brennan who provided us with guidance where needed, and, who, as the Director of CASMA, provided us with a home to conduct this work. Thanks to Jennifer Jones for her secretarial 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 College Board Vice President Wayne Camara for his substantial support of this project as well as College Board staff Amy Hendrickson, Gerald Melican, Rosemary Reshetar, and Kevin Sweeney, and former College Board staff member Kristen Huff.

Michael J. Kolen
Won-Chan Lee
December, 2012
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.

Contents

Chapter 1: Introduction and Overview for Volume 2

Michael J. Kolen and Won-Chan Lee

The University of Iowa, Iowa City, IA

Abstract

This chapter provides an overview and describes methodology that is common to the research studies on psychometric methods for mixed-format tests that are reported in this volume. This chapter highlights some of the major methodological issues encountered and some of the major findings. In addition, it relates the findings from this Volume 2 to Volume 1. The discussion section provides a summary of some of the major findings that can be used to inform best practices in equating alternate forms of mixed-format tests.

Introduction and Overview for Volume 2

The research described in Volume 2 is closely related to the research conducted in Volume 1 (Kolen & Lee, 2011). This chapter provides an overview of Volume 2. It describes methodology that is common across chapters of Volume 1 and Volume 2. In addition, this chapter highlights some of the major methodological issues encountered and some of the major findings across both volumes.

Although all of 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. Refer to Volume 1 for more detail on the data and some of the equating methods that are investigated in Volume 2.

The chapter begins with a description of the research questions, designs, and findings. The findings from Volume 2 are related to findings in Volume 1. The chapter concludes with a summary of some of the important practical findings from the research reported in Volume 1 and Volume 2.

Research Questions, Designs, and Findings

The chapters in this monograph differ from one another and from the chapters in Volume 1 in the research questions addressed and the designs used to address them. Chapters 2 through 4 consider various issues associated with equating of mixed-format tests using the common item nonequivalent groups design. Chapter 5 proposes an IRT observed-score equating procedure under the simple-structure multidimensional IRT framework for a random groups equating design. Chapter 6 investigates item response theory (IRT) item fit statistics. The research questions, designs, and findings are summarized in this section.

Chapter 2

Chapter 2 addressed the conditions under which multiple-choice (MC) items can be used to adequately equate mixed-format tests using a set of common items that consists solely of MC items. Data were simulated from a simple structure multidimensional IRT (SS-MIRT) model where the MC items were assumed to assess a different, but correlated, unidimensional proficiency from the free-response (FR) items. Data from the AP World History exam were fit using the SS-MIRT model, and the parameter estimates were used as parameters for the simulation. The correlation between the MC and FR constructs and the extent of difference in average proficiency of the examinees taking the forms being equated were varied in the simulation. Equipercentile methods were used in this study, and the Difference That Matters (DTM) for raw scores, scale scores, and AP grades were used to provide a practical criterion for considering the results.

The results suggested that when using only MC items as common items for equating mixed-format tests, equating error increased as the correlation between the constructs assessed by the MC and FR items decreased. The DTM criterion was used to provide practical guidance as to how large the correlation needs to be for accurate equating.

Chapter 3

The effects of group differences on equating results for mixed-format tests were investigated in Chapter 3. The study compared the influence of group differences on equating results for both equipercentile and IRT equating methods, and examined results from three different AP exams that had different magnitudes of true-score correlations between MC and FR items.

The study considered two different types of criteria for comparing the equating results based on real test data. For one criterion, operational test forms were equated. For this criterion, pseudo groups of examinees were formed by oversampling students based on demographic characteristics. The pseudo groups were formed such that the groups taking the old and new forms differed in average proficiency by varying amounts in effect size (ES) units (0.0, 0.2, and 0.4). The equating for the ES of 0.0 was used as the criterion equating for each equating method studied. A strength of this design and criterion is that intact operational forms were equated. However, this criterion might be questioned because the groups were formed based on sampling using demographic characteristics.

The second criterion was based on using pseudo-test forms. A pair of pseudo-test forms and a set of common items were constructed from a single test form. Using examinee data on that test form, the pseudo-test forms were equated using the single group design. The single group design equipercentile equating was used as the criterion for the equipercentile methods. Single group IRT true and observed score equatings were used as the criteria for the IRT equatings. As was done with the first criterion, pseudo groups were formed to differ by ES units of 0.0, 0.2, 0.4, and 0.6. A strength of this criterion is that a single group equating, which would be expected to be stable, is used as the criterion equating. A weakness is that the pseudo-test forms might not well represent operational test forms (for example, they are much shorter).

Common item nonequivalent equipercentile and IRT equating results were compared using both criteria. One focus of this study was on whether the same general results were found for the two criteria.

Equating tended to be more accurate (less biased) when there were smaller group differences. With larger group differences, chained equipercentile (CE) equating and IRT equating were found to be more accurate (less biased) than frequency estimation (FE) equipercentile equating. CE tended to have larger standard errors of equating than the other methods. Findings for the two criteria were found to be similar.

Chapter 4

The usefulness of matched samples equating methods was investigated in Chapter 4 using pseudo groups that were formed by oversampling students based on the educational level of the examinee’s parents so that the groups taking the old and new forms differed in average proficiency by varying amounts in effect size (ES) units (0.0, 0.25, 0.50, and 0.75). Pseudo-test forms were created and scores on these forms were equated using common-item nonequivalent groups equipercentile and IRT methods.

Prior to equating, examinee groups taking the old and new forms were either not matched, matched on educational level of the examinee’s parents (the variable used for selection), matched on propensity scores based on demographic variables that included educational level of the examinee’s parents, or matched on propensity scores based on demographic variables that did not include the educational level of the examinee’s parents. The equating relationships for the equating when ES=0.0 were used as the criterion equatings. The effect of group differences on equating assumptions was evaluated for the equating methods.

As group differences increased, the estimated equating relationships became more biased. Matching that included the variable used to select the examinee groups (i.e., educational level of the examinee’s parents) led to more accurate equating compared to matching only on other variables or on no matching. FE was most sensitive to group differences and benefited the most by matching. Matching that included the variable used to select examinee groups greatly improved the degree to which the assumptions held for the FE and CE methods.

Chapter 5

An SS-MIRT equating procedure was developed in Chapter 5. As in Chapter 2, for the SS-MIRT equating procedure, the MC items were assumed to assess a different, but correlated, unidimensional proficiency from the FR items. The SS-MIRT equating procedure and a unidimensional IRT model equating procedure were used to equate AP World History forms. In addition, using the fitted SS-MIRT model as a base, test data were generated for which the correlation between the MC and FR proficiencies were 0.5, 0.8, and 0.95. In this simulation, the equating results for the SS-MIRT and unidimensional procedures were compared. The SS-MIRT method was found to produce adequate equating and outperformed the unidimensional IRT equating methods when the correlation between the MC and FR proficiencies was less than perfect.

Chapter 6

Different IRT item fit statistics for mixed-format tests were compared in Chapter 6. These statistics can be used to help screen items for fit to a unidimensional IRT model. In this study, data on 13 different AP test forms were used to study item fit statistics. Item fit statistics were calculated and compared. Item fit statistics, sample size, IRT model used for the FR items, and the minimum expected cell frequency used were varied in this study.

All of the item fit statistics identified more items as misfitting as sample size increased. The statistics proposed by Orlando and Thissen (2000) were less affected by sample size than the other statistics and tended to identify fewer items as misfitting than the other statistics studied. Statistics developed using similar approaches agreed more with one another in terms of items identified as misfitting. The two different IRT models used for the FR items produced similar results. One of the conclusions from this study is that item fit statistics that are not sample size dependent are preferable.

Discussion and Conclusions

The research questions for the four Volume 2 chapters that examined the equating of mixed-format test forms are given in Table 1. As can be seen, the research questions address a series of practical issues in equating, including the use of MC and FR items in common item sets, the influence of group differences and matched sampling on equating results for various methods, and the impact of equating design, method and evaluation criteria on equating results. Along with the research questions addressed in Volume 1, a wide variety of research questions regarding applied issues in equating of mixed-format tests have been considered in these two volumes.

Many of the studies in Volumes 1 and 2 examined issues in common-item nonequivalent groups equating for mixed-format tests using real data. The design characteristics of these studies are shown in Table 2. As is evident from this table, pseudo-test forms and pseudo groups were used in many of these studies. The use of pseudo groups, where assignment to group is based on background variables, is a useful way to study the effect of the magnitude of group differences on equating results. Note that scores on the tests to be equated or on the common items were NOT used to form these groups. Doing so would have resulted in a situation in which group membership would be correlated with measurement error. Basing assignment on variables other than the test scores avoids this correlation.

Some of these studies also made substantial use of pseudo-test forms, which permits a single group equating relationship to serve as a criterion for comparing equating results. This criterion is a very convenient and useful.

Chapter 2 of Volume 2 used a simulation to assess the effects of the correlation between the MC and FR proficiencies and magnitude of group differences on the adequacy of mixed-format equating using only MC items as common items. The use of simulation procedures allowed for substantial control over the correlation and group difference conditions. Simulation also did not require the intensive process of sampling items and persons from existing operational data to form pseudo-test forms and pseudo groups.

Chapter 2 of Volume 1 used intact forms and intact groups to compare the similarity of equating results and standard errors of equating for different equating methods. The strength of the use of intact groups is that the groups differ in realistic ways and the results pertain to intact, operational test forms.

Each of the types of studies has drawbacks and limitations. With intact forms and intact groups, it is not possible to estimate equating bias or to assess which equating procedure produces results with the least amount of overall error. With pseudo groups, the group differences observed might differ in important ways from the group differences that occur in operational equating. With pseudo-test forms, the forms that are studied are not intact forms so it is difficult to tell whether the results pertain to intact operational forms. Simulation studies are only useful to the extent that the model used for the simulation is reasonably consistent with reality. Therefore, it appears that each type of study is important, and that none provides definitive answers by itself. A strategy that involves using multiple designs and criteria for equating research studies seems necessary to inform best practices. When the findings from different designs and criteria converge (as they do, for example, in Chapter 3 of Volume 2), the reasonableness of those findings is more compelling.

Some of the important practical findings from the research reported in Volumes 1 and 2 are as follows: First, it appears that the FE method provided less accurate equating than the CE or IRT methods when group differences were large. However, the FE method tended to have less random error than CE and less overall error when group differences were small (e.g., .05 ES units or less). Second, equating was more accurate when group differences were small. Third, presmoothing and postsmoothing improved equipercentile equating precision. Fourth, MC-only sets of common items appeared to provide reasonable equating results when the correlation between MC and FR proficiencies were high and the examinee group differences in proficiency were not large. Fifth, matched samples equating can improve equating accuracy, but it is unclear whether or not the variable or set of variables that underlie group differences would be known in practice. Sixth, the SS-MIRT method appeared promising as a method for equating mixed-format tests when the correlation between the MC and FR proficiencies were substantially less than 1. Seventh, new IRT item fit criteria should be developed that are less sensitive to sample size.

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.

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

Table 1

Highlights of Research Questions Chapters 2-5 in Volume 2

Chapter	Research Questions
2	Under what correlations between MC and CR proficiencies and examinee group differences can only MC items be used as common items to adequately equate mixed-format tests?
3	(a) To what extent is the magnitude of random and systematic equating errors differentially related to group differences for various equating methods? (b) To what extent do research findings using two different research designs and criteria lead to similar findings?
4	Does matched samples equating improve equating accuracy?
5	Does a simple structure multidimensional IRT equating procedure improve equating accuracy compared to a undimensional IRT model when MC and FR items measure different proficiencies for a mixed-format test?

Table 2

Highlights of Design Characteristics for Studies in Volume 1 Chapters 2 through 6 and Volume 2 Chapters 3 and 4

Design Characteristic	Vol. 1 Ch. 2	Vol. 1 Ch. 3	Vol. 1 Ch. 4	Vol. 1 Ch. 5	Vol. 1 Ch. 6	Vol. 2 Ch. 3	Vol. 2 Ch. 4
Examinee Groups	Intact	Pseudo	Pseudo	Pseudo	Pseudo	Pseudo	Pseudo
Variable Used to Form Pseudo Groups	None	Reduced Fee Indicator	Gender	Ethnicity and Parental Education	Parental Education	Ethnicity and Parental Education	Parental Education
Test Forms	Intact	Intact	Pseudo	Pseudo	Pseudo	(a) Pseudo (b) Intact	Pseudo
Used Resampling	Yes	No	Yes	Yes	Yes	Yes	Yes
Criterion Equating(s)	None	Total Group	Single Group	Single Group	(a) Single Group (b) No Group Differences CINEG	(a) Single Group (b) No Group Differences CINEG	No Group Differences CINEG
Common Item Composition	MC	MC	MC	MC and MC/FR	MC	MC and MC/FR	MC

Chapter 2: Equating Mixed-Format Tests Using Dichotomous Common Items

Won-Chan Lee, Yi He, Sarah Hagge, Wei Wang, and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

The main purpose of this paper is to evaluate the feasibility of equating mixed-format test forms using a common-item set that consists solely of multiple-choice (MC) items. More specifically, this study investigates the effects of (a) the correlation between the two constructs measured by MC and free response (FR) items, and (b) group ability differences on equating for mixed-format test forms. A series of simulation studies is conducted. Six traditional equating methods are considered. Consistent with previous research, the results of this study show that the magnitude of the error increases as the correlation between the two constructs decreases or as the group difference increases. In general, a higher correlation is needed as the group difference increases to achieve adequate equating. A minimum level of the correlation necessary for adequate equating is determined for each combination of simulation conditions based on the comparison of overall and conditional bias statistics to a Difference That Matters value.

Equating Mixed-Format Tests Using Dichotomous Common Items

Free response (FR) items are often claimed to provide a more direct measure of examinees' high level thinking and reasoning skills than multiple-choice (MC) items (Lane & Stone, 2006). Mixed-format tests consisting of both MC and FR items have become more prevalent in many recent testing programs because of the view that mixed-format tests can achieve dual goals of high reliability and measurement of complex skills. Typically, the number of FR items administered to each examinee tends to be very small, and the tasks are memorable. One of the vexing problems with equating for mixed-format tests under the common-item nonequivalent groups design is that the issue of item security often prevents a set of FR items that were used in a previous administration from being included in a new form of a mixed-format test. There are also concerns with the comparability of judges' ratings from one test administration to the next with the FR common items. Therefore, common items in alternate forms of a mixed-format test typically are comprised of only MC items, which can be of a concern from the perspectives of representativeness of common items to the total test.

Several studies have been conducted to explore the effectiveness of using MC items only as a common-item set in linking and equating for mixed-format tests. In these studies, the following were identified as important factors: (a) the correlation between MC and FR scores (Kirkpatrick, 2005; Paek & Kim, 2007; Walker & Kim, 2009), (b) the correlation between the common-item scores and entire test scores (Wu, Huang, Huh, & Harris, 2009), (c) the ratio of MC to FR score points (Paek & Kim, 2007; Tan, Kim, Paek, & Xiang, 2009), (d) when the group differences in difficulty for the CR items are similar in magnitude to the group differences on the MC items (Hagge & Kolen, 2011), and (e) group differences (Kirkpatrick, 2005; Cao, 2008). In general, better equating results were associated with a high correlation between MC and FR scores, a high correlation between common-item scores and total test scores, a large ratio of MC to FR score points, and a small group difference.

The main purpose of this paper is to evaluate the feasibility of equating mixed-format test forms using a common-item set that consists solely of MC items. More specifically, this study investigated (a) how large the correlation between the two constructs measured by MC items and FR items needs to be to do equating for mixed-format test forms; and (b) how small group differences need to be to do equating for mixed-format test forms. A series of simulation studies was conducted to address these issues.

Method

Two forms (old and new) of the Advanced Placement (AP) World History test were used as the basis for the simulation. Each form had 70 MC items and 3 FR items (scored 0-9). Operationally, composite scores on the test are weighted sums of formula scores on the MC and FR sections. The weights for the MC and FR sections for this test were non-integer values of approximately .86 and 2.2, respectively. The common-item set consisted of 22 MC items, and the common-item scores contributed to the total test scores (i.e., internal anchor).

Several modifications were considered for the current simulation study, which made the simulation data somewhat remote from the operational characteristics of the test. First, to avoid complexities in use of IRT, examinee records were eliminated from the data if there were any omitted item responses. The resulting sample sizes for the old and new form data were 5,962 and 6,136, respectively. Second, number-correct scoring, instead of formula scoring, was used for the MC items. Lastly, the non-integer weight for each section was rounded to the nearest integer, which resulted in integer weights of 1 for the MC section and 2 for the FR section. In this weighting scheme, the relative contributions of the MC and FR sections to the composite score ranging from 0 to 124 in terms of possible score points were 56% and 44%, respectively. The MC common-item score contributed about 18% to the composite score again, in terms of possible score points. The changes made to AP World History data in this simulation greatly simplified the simulation process and are also relevant to address the research questions of the present study. However, these changes reduce the generalizability of the results to World History.

The three parameter logistic model (Lord, 1980) and the graded response model (Samejima, 1997) were used to fit the MC and FR items, respectively. Calibration was done separately for each form and each of the MC and FR sections using PARSCALE (Muraki & Bock, 2003). Separate calibrations for the old and new form data resulted in two sets of item parameter estimates for the common items. The parameter estimates for the common items in the new form were replaced with those for the common items in the old form such that there was only one set of item parameter estimates for the common-item set. These estimated item parameters were treated as generating (or “true”) item parameters for the old and new forms.

True Equating Relationships

The population (or true) equating relationships were established by conducting simple-structure multidimensional IRT observed-score equating (Lee & Brossman, 2012) based on the generating item parameters and a population bivariate normal distribution for the two latent variables. Let $BN(\mu_{MC}, \mu_{FR}, \sigma_{MC}, \sigma_{FR}, \rho)$ denote a bivariate normal distribution with means, standard deviations, and correlation for the two θ distributions for MC and FR sections. It was assumed that the MC items measured only one construct, the FR items measured another construct, and the two constructs were allowed to be correlated. In the terminology of factor analysis, this is a simple structure model. The bivariate normal distributions were identified in accordance with the simulation factors of the correlation levels and population means. Each simulation condition involved a pair of bivariate population distributions, one for the old form group and the other for the new form group. For each of the old and new form groups, 41x41 pairs (the MC and FR make up the pair) of theta quadrature points and weights were obtained for the specified population distribution, with theta values ranging from -4 to 4 for each of the MC and FR sections.

For each of the old and new form populations, a fitted marginal observed-score distribution was obtained using the bivariate normal quadrature distribution and the generating item parameters. Based on the two fitted observed score distributions for the new and old form populations, equipercentile equating was conducted to obtain new-form equated scores. (The old-form conversion table used for equating is discussed later.) More specifically, conditional raw-score distributions were first obtained using an extended version of the Lord-Wingersky (Lord & Wingersky, 1984) algorithm (Hanson, 1994; Thissen, Pommerich, Billeaud, & Williams, 1995) by loading the MC items on the MC dimension and the FR items on the FR dimension—that is, for each pair of theta quadrature points, category probabilities for each item were computed using the item's true parameters and the theta value associated with the item's dimension. Then, the recursive algorithm was used to compute the raw-score distribution for a given pair of theta values. These conditional raw-score distributions were aggregated over the entire bivariate theta distribution to obtain the marginal observed-score distribution. Equipercentile equating was conducted on an equally-weighted synthetic population of the old and new form populations. Even though the same generating item parameters were used, the true equating relationships were different for different simulation conditions of population distributions.

Simulation Factors and Procedure

The simulation factors considered in the present study were:

- 11 levels of correlation (.5 to 1.0 with an increment of .05) between the latent constructs measured by MC and FR items;
- five levels of ability effect-size difference between the new and old form groups (*ES* = .05, .1, .2, .3, or .5);
- one level of sample size (*N* = 3000);
- three types of score scales (raw composite scores, normalized scale scores, and AP grades); and
- six equating methods.

A total of 55 (11x5) simulation conditions were studied and results were obtained for each of six equating methods and each type of score scale. As discussed later, additional simulations were carried out to examine the performance of the cubic spline postsmoothing procedure with different smoothing parameter values. The levels of latent-trait correlation were chosen according to estimated disattenuated correlations of the MC and FR scores for other AP exams, which typically range from .75 to 1.0. The lowest correlation value was set at .5 in this study to establish a wider coverage. The five levels of effect size were determined based on preliminary real data analyses using various AP exams. The effect sizes based on examinees' scores on common-item sets ranged from 0.01 to 0.36. Only one level of sample size was employed.

The six equating methods (Kolen & Brennan, 2004) were: unsmoothed frequency estimation (UnSm_FE), frequency estimation with cubic-spline postsmoothing (PostSm_FE), frequency estimation with log-linear presmoothing (PreSm_FE), unsmoothed chained equipercentile (UnSm_CE), chained equipercentile with cubic-spline postsmoothing (PostSm_CE), and chained equipercentile with log-linear presmoothing (PreSm_CE). The smoothing parameter for cubic-spline postsmoothing was .1, which was selected by examination of equating results based on the original data. A bootstrap resampling procedure with 1000 bootstrap replications was employed to estimate the standard error of equating for the cubic-spline postsmoothing method. For log-linear presmoothing, the smoothing parameters were 6 for the marginals and 1 for the cross product. Synthetic population weights of .5 for both populations were employed for the frequency estimation method.

Five different levels of ability effect size were considered:

- ES = .05: $(\theta_{MC}, \theta_{FR})_{Old} \sim BN(0,\ 0,\ 1,\ 1,\ \rho)$ and $(\theta_{MC}, \theta_{FR})_{New} \sim BN(.05,\ .05,\ 1,\ 1,\ \rho)$
- ES = .1: $(\theta_{MC}, \theta_{FR})_{Old} \sim BN(0,\ 0,\ 1,\ 1,\ \rho)$ and $(\theta_{MC}, \theta_{FR})_{New} \sim BN(.1,\ .1,\ 1,\ 1,\ \rho)$
- ES = .2: $(\theta_{MC}, \theta_{FR})_{Old} \sim BN(0,\ 0,\ 1,\ 1,\ \rho)$ and $(\theta_{MC}, \theta_{FR})_{New} \sim BN(.2,\ .2,\ 1,\ 1,\ \rho)$
- ES = .3: $(\theta_{MC}, \theta_{FR})_{Old} \sim BN(0,\ 0,\ 1,\ 1,\ \rho)$ and $(\theta_{MC}, \theta_{FR})_{New} \sim BN(.3,\ .3,\ 1,\ 1,\ \rho)$
- ES = .5: $(\theta_{MC}, \theta_{FR})_{Old} \sim BN(0,\ 0,\ 1,\ 1,\ \rho)$ and $(\theta_{MC}, \theta_{FR})_{New} \sim BN(.5,\ .5,\ 1,\ 1,\ \rho)$

Note that the population means for the MC and FR sections for the old form group were always zero and the standard deviations were all one. The five levels of the effect size represented the group mean differences between the two groups. For each of the five effect size conditions, 11 different levels of correlation from .5 to 1.0 with an increment of .05 were examined. A lower correlation value indicates that the two latent constructs each measured by the MC and FR items, respectively, are more dissimilar and thus equating with the MC-only anchor test presumably is more problematic. The correlation value for each condition was the same for the old and new form populations.

The scores that are reported to examinees for actual AP exams are 1-5 integer grades. It is likely that this score scale is too coarse to reveal important differences in the equated scale scores for different equating methods. The raw composite scores (simply referred to as raw scores hereafter) also have the potential problem of overemphasizing small differences in equated scores if the number of score points in the reported score scale such as the AP grades is much smaller than the number of raw-score points. A normalized score scale was constructed using a method described by Kolen and Brennan (2004), which had a lot more score points than the AP grades but less than the raw scores. The normalized score scale ranging from 0 to 70 had a mean of 35 and standard deviation of 10. Equating results were obtained for three different types of score scales: raw scores, normalized scale scores, and AP grades. A conversion table was created for the old form, in which raw scores ranging from 0 to 124 converted to 0-70 unrounded normalized scale scores and 1-5 integer grades. Grades were assigned such that the percent earning each grade was similar to that for this group for the operational AP grades. The same old-form conversion tables were used for all simulation conditions. Table 1 provides the old-form conversion table that was used in this simulation study. Note that unrounded values were used for the normalized score scale, while rounded integer values were used for the AP grades. Equating results were evaluated by comparing the new-form true and estimated equating relationships in terms of unrounded equated scores for all three types of score scales. Evaluating unrounded equated scores, as opposed to rounded scores, is more consistent with the use of non-integer Differences That Matter statistics as discussed later.

For each condition of population distributions, the following simulation steps were used:

1. Draw three thousand (*N* = 3,000) bivariate normal deviates (i.e., pairs of theta values) from each of the old and new form populations. Each pair of theta values represents an examinee.
2. Generate item responses for each form using the generating item parameters and the examinees' theta values. The three-parameter logistic and graded response models were used for the MC and FR items, respectively.
3. Conduct equating based on the simulated data to obtain new-form unrounded equated scale scores using the six equating methods considered in this study.
4. Repeat the above steps 100 times.

For each replication, each equating method produced a new-form conversion table containing: a) new-form integer raw scores, b) unrounded equated raw scores, c) unrounded equated normalized scale scores, and d) unrounded equated grades. Over 100 replications, 100 new-form conversion tables were produced for each equating method.

Summary Statistics

The equating results for each of the six equating methods over 100 replications were summarized by mean squared error (MSE), squared bias (SB), and variance (VAR). These statistics were computed for each raw-score point (i.e., conditional statistics) and across all score points (i.e., overall statistics). The conditional SB was computed as:

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

where e_x is the true equated score at raw score *x*; $\hat{e}_{xr}$ is an estimated equated score at raw score *x* on replication *r*; and $(1/100)\sum_{r=1}^{100} \hat{e}_{xr}$ is the mean, over replications, of estimated equated scores at a given raw score. The conditional variance was given by

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

The conditional MSE was computed as the sum of conditional SB and conditional variance: $MSE(x) = SB(x) + VAR(x)$. The corresponding overall statistics across all score points were computed as weighted averages of the conditional statistics where the weights were the frequencies of new-form scores based on the IRT model. The overall statistics were weighted by frequencies, as opposed to using equal or uniform weights, for the purpose of alleviating the influence of equating results at very high or low scores where almost no data existed. In addition to the weighted overall statistics, graphical representation of the conditional statistics that vary across the score scale will be provided to show the results that do not involve weighting. Both the conditional and overall statistics were computed for all three types of scale scores.

These summary statistics can provide a standard to evaluate the relative performance of the six equating methods; however they do not answer the question of how small the magnitude of error should be for adequate equating. To determine an "acceptable" level of error, the notion of Differences That Matter (DTM) was used. The DTM was defined as .5 for all three types of scale scores in the present study according to the rationale that a difference in unrounded scale scores that is greater than .5 can alter rounded reported scale scores (Dorans & Feigenbaum, 1994). The DTM was compared to $\sqrt{SB(x)}$. When $\sqrt{SB(x)}$ was greater than the DTM, the equating was considered unacceptable. Note that the DTM was compared with the SB rather than the MSE because both the DTM and the SB, in a sense, reflect the "differences" between the true and estimated equating relationships. Note also that the DTM value of .5 used in this study should be viewed as a practically convenient benchmark rather than an absolute criterion.

Disattenuated Correlations

In this simulation study, the correlation between the two constructs measured by the MC and FR items was expressed in the IRT theta metric, which is different from the disattenuated correlation under classical test theory. Given that the classical disattenuated correlation between the MC and FR summed scores typically is used in practice to determine the degree to which the two item-format sections are correlated, the classical disattenuated correlation was computed for each sample of each condition to examine whether the true theta correlation was reasonably similar to the classical disattenuated correlation. Note that the term “classical” indicates that coefficient alpha was used as a reliability estimate.

Results

Estimated Correlations

The calculated classical disattenuated correlations across all the simulation conditions and all the samples were surprisingly close to the true theta correlations. Across all conditions investigated in this study, the classical disattenuated correlation was always lower than the true theta correlation, but only by a value in the range of .005 to .01. This finding clearly indicates that the use of theta correlations in this simulation study can provide adequate information when the classical disattenuated correlation is used in practice to identify the degree of association between the MC and FR sections. Note that the remarkably similar values between the latent theta correlations and classical disattenuated correlations might be due to the fact the test characteristic curves that relate the IRT theta values and the classical true scores are almost linear across a wide range of the theta scale.

Raw Scores

Tables 2 through 6 display the raw-score summary statistics (i.e., overall MSE, SB, and VAR) for the five effect size conditions, respectively. The results for the six equating methods for the 11 correlation conditions are reported in the tables. (As discussed later, the shaded cells in the squared bias portion of the tables indicate that the equating results under those conditions were adequate relative to the DTM.)

In general, the magnitude of the MSE tended to increase as the correlation decreased primarily due to a tendency for increase in the SB. Also, the comparison across the five tables showed that the MSE tended to be larger, due to the larger SB, when the effect size was greater. When the effect size was small (i.e., .05 or .1), the main source of error was the VAR rather than the SB. By contrast, when the effect size was larger (i.e., .2 or greater), the main source of error was the SB.

The results for the unsmoothed equating methods tended to show slightly larger MSE values, mainly due to the larger VAR, than those for the presmoothed and postsmoothed equating methods for almost all correlation levels and effect sizes. This finding is consistent with the conclusion that smoothing procedures help reduce sampling error.

Comparing the FE and CE methods, the FE methods produced smaller MSE values than the CE counterparts when the effect size was very small (i.e., Table 2), except for the slightly larger MSE for the presmoothed FE than the presmoothed CE when the correlation was very low. The smaller MSE for the FE methods compared to the CE methods in Table 2 was mainly attributable to the smaller VAR; the SB for the FE methods was almost always larger than that for the CE methods. When the effect size was greater than .05, (i.e., Tables 3 through 6), the MSE and SB for the FE methods were substantially larger than those for the CE methods. Across all conditions, however, the FE methods had smaller VAR than the CE methods.

The comparison between the postsmoothed and presmoothed equating methods indicates that the MSE values were similar, yet slightly smaller for the presmoothed methods; the SB tended to be larger for the presmoothed methods; and the VAR was larger for the postsmoothed methods. When $ES = .5$, however, the postsmoothed methods showed smaller MSE values. The relative performance of the two smoothing procedures may well depend upon the choice of parameter values for smoothing.

To further address this issue, an alternative smoothing parameter value of *S* = .3, instead of .1, was considered for the postsmoothing procedure and a representative set of simulation conditions were replicated. A larger parameter for the cubic spline postsmoothing procedure results in a smoother fitted distribution of the equivalents. Table 7 summarizes the raw-score results for this additional simulation. Results are provided for the postsmoothed FE and CE methods using *S* = .1, the presmoothed methods, and the postsmoothed methods using *S* = .3. The top panel in the table shows results for $ES = .05$; the middle panel displays results for $ES = .2$; and the bottom panel exhibits those for $ES = .5$. Across all the conditions of three correlation levels (.5, .8, and 1.0) and three effect size levels, the MSE values associated with the postsmoothed methods with *S* = .3 were smaller than those for the presmoothed methods and the postsmoothed methods with *S* = .1. This is noteworthy given that *S* = .3, compared to *S* = .1, reduced even the amount of bias for the FE methods in many conditions, which was a somewhat unexpected result because the use of a higher parameter value is likely to reduce the variance but may increase the bias. Both the bias and variance tended to be lower for the postsmoothed methods with *S* = .3 than for the presmoothed methods.

In order to determine an “acceptable” level of equating, the SB values in Tables 2 through 6 were compared to the squared DTM value of .25. Any SB value lower than .25 was highlighted in the tables to indicate that the equating results are acceptable for a particular condition. Overall, it was obvious that as the effect size increased a higher correlation was needed to obtain acceptable equating results. As shown in Table 2, when the effect size was very small (i.e., $ES = .05$), all six equating methods produced results that were acceptable across all correlation levels. When the effect size was .1, Table 3 shows that all CE methods (i.e., unsmoothed, presmoothed, and postsmoothed) had acceptable equating results for all correlation levels, while all FE methods showed acceptable equating results if the correlation was .9 or higher. For the condition of $ES = .2$ as displayed in Table 4, the CE methods produced acceptable equating results when the correlation was .8 or higher, and none of the FE methods were acceptable for any correlation level. When the effect size was either .3 or .5 (see Tables 5 and 6), no equating method demonstrated an acceptable level of bias, except for the CE methods under the condition of $ES = .3$ and ρ = 1.0.

To evaluate the amount of bias in the equating results across different score levels, conditional bias was compared to the DTM for all six equating methods in Figures 1 through 4. Figure 1 displays the results for $ES = .05$; Figure 2 shows the results for $ES = .1$; Figure 3 for $ES = .2$; and Figure 4 for $ES = .3$ and $ES = .5$. Each of the plots in the figures for a particular effect size condition represents a different correlation condition. (Note that results for a selected set of correlation levels are provided in each figure.) In order to compare results for difference types of score scales (as discussed later), both the DTM of .5 and $\sqrt{SB(x)}$ for raw scores were standardized by dividing them each by the standard deviation of the raw-score distribution which was computed based on the true raw-score equating relationships and the new form population. A straight line in the middle of each plot represents the standardized DTM, and the six curvy lines are the bias for the six equating methods, respectively. Notice that the horizontal axis representing the old-form raw scores was truncated at both tails where there were insufficient data to show meaningful equating results.

Focusing on Figure 1, the amount of bias for all six equating methods generally was below the straight DTM line across the most range of the raw-score scale for all correlation levels, including the lowest correlation of .5. This finding is consistent with the general observations made from the overall summary statistics.

Displayed in Figure 2 are the conditional standardized bias and DTM for the condition of $ES = .1$. As the effect size increased the bias lines tended to cluster into two groups. The three lines clustered together with higher bias values were for the three FE methods, and the other cluster of three lines with lower bias values represented the three CE methods. The requirement of at least ρ = .9 for the FE methods based on the overall bias statistics seemed to be well supported by the conditional results. However, the conclusion that the CE methods based on the overall statistics worked adequately even if the correlation was as low as .5 might need to be altered. The bias of all CE methods with relatively low correlations tended to exceed the DTM near high scores and the presmoothed CE showed the large bias near the score point of 60. Therefore, it would be safe, although somewhat conservative, to require a correlation of at least .7 for the CE methods when $ES = .1$.

The plots presented in Figure 3 for $ES = .2$ showed a more evident clustering pattern for the CE and FE methods. The same conclusions as those based on the overall statistics seem reasonable; namely, any of the FE methods did not provide adequate equating even under the perfect correlation condition, while the bias for the three CE methods tended to be smaller than the DTM when the correlation was higher than .8.

When $ES = .3$, the CE methods seemed to provide adequate equating only if the correlation was 1.0; while when $ES = .5$, the conditional bias for all six equating methods tended to exceed the DTM criterion across all correlation conditions even under the condition of a perfect correlation. Figure 4 provides the results for $ES = .3$ and $ES = .5$ for the perfect correlation condition.

Normalized Scale Scores

As with the raw scores, the results for the normalized scale scores were evaluated by both overall summary statistics and conditional bias plots relative to the DTM value of .5. Although tables and figures are not provided, important findings are summarized here. Because of the smaller number of score points in the normalized score scale compared to the number of the raw-score points, the magnitudes of the statistics were much smaller than those for the raw-score results. As a result, more correlation conditions turned out to be acceptable. The relative performances of the six equating methods were very similar to those observed based on the raw-score results and thus not further discussed here.

The overall summary statistics suggested that the results for all equating methods were acceptable based on the DTM criterion when $ES = .05$ or $ES = .1$. When $ES = .2$, the results for all CE methods were acceptable, but equating for the FE methods was adequate only if the correlation was .85 or higher. When $ES = .3$, no equating was acceptable for the FE methods, while the CE methods showed acceptable results if the correlation was .7 or higher. Finally, when $ES = .5$, results were acceptable only if the CE methods were used and the correlation was one. The inspection of plots for the standardized conditional bias against the standardized DTM suggested retaining the conclusions based on the overall summary statistics, except for one condition—when $ES = .5$, a slight modification of the previous conclusion was made to require .9 or higher rather than 1.0 for the CE methods.

AP Grades

With only five score points in AP grade levels, the error statistics were remarkably smaller than those for the normalized scale scores and raw scores. Nonetheless, the comparison of the various equating methods in their relative performance led to the same conclusions as before. The overall SB values compared to the squared DTM of .25 suggested that all equating results for all correlation levels and all effect size conditions were acceptable. This somewhat extreme result can be explained by the fact that, with only five score points, most raw-score points converted to the same equated grade level over replications and there existed only a small number of raw-score points near the 1/2, 2/3, 3/4, and 4/5 grade cut scores which had a higher probability of being associated with different equated grade levels over replications. Consequently, a substantial amount of error was concentrated on those few raw-score points and, by contrast, relatively much less error was generated for the rest of the raw-score points. For this reason, examination of conditional bias was important.

Consistent with the results of the overall SB, the conditional results showed that the bias for all six equating methods was lower than the DTM across all correlation conditions when $ES = .05$. For the $ES = .1$ condition, all CE methods had an acceptable level of the bias regardless of the correlation levels, while the bias of the FE methods appeared to be lower than the DTM if the correlation was higher than .7, except for the presmoothed FE method, for which a perfect correlation was needed to achieve an acceptable level of the bias. When $ES = .2$, all the CE methods provided adequate equating even with the lowest correlation condition, while the FE methods did not show an acceptable level of the bias for any correlation level. None of the FE methods worked adequately, when $ES = .3$; while the CE methods produced an acceptable level of the bias when the correlation was .85 or higher. Lastly, the conditional bias of the all six equating methods, when $ES = .5$, exceeded the criterion across all correlation levels.

Summary of Results

Considering both the overall SB and conditional bias together with the DTM as a criterion, minimum correlation levels to achieve adequate equating results for the two groups of equating methods (FE and CE), five different effect sizes, and three different types of scale scores are summarized in Table 8. Decisions were made by identifying a correlation level, for which the equating results meet the DTM criterion for the overall statistics and across most of the range of the score scale. Note that the six equating methods were grouped into either FE or CE because the three versions within each group (i.e., unsmoothed, presmoothed, and postsmoothed) tended to perform similarly, although there were a few exceptions. Given an effect size, “A” means that the results are acceptable across all correlation levels with the lowest being .5; “N” means that the results are not acceptable even with a perfect correlation condition.

Summary and Discussion

For a common-item nonequivalent groups design, having a common-item set that is a mini version of the total test is conducive to adequate equating (Kolen & Brennan, 2004). When a test consists of both MC and FR items and equating is done using the common item nonequivalent groups design, it is often the case that a common-item set (i.e., anchor items) is comprised of MC items only, due to various practical and psychometric reasons. This obviously violates the “mini-test” property of the common-item set. It is critical to study under what conditions this type of equating will or will not be adequate. The present study is an attempt to address this question.

The classical disattenuated correlation appears to be a useful statistic to approximate the degree to which the MC and FR dimensions are correlated on the IRT ability metric. An estimated disattenuated correlation and an estimate of the effect size could be used to provide some practical guidelines about the adequacy of this type of equating. The effect size, for example, can be approximated using the performance of the old and new form groups on the common items.

The results of this study reveal that the magnitude of error increases as the correlation between the MC and FR latent variables decreases. The error also increases as the effect size of the group ability differences increases. These results are consistent with findings from previous studies (Cao, 2008; Hagge & Kolen, 2011; Kirkpatrick, 2005; Paek & Kim, 2007; Walker & Kim, 2009). Overall, it seems that a higher correlation is needed as the effect size increases to achieve adequate equating with the DTM criterion. The result of larger error for the FE method compared to the CE method, when the effect size is large, is consistent with the results reported in Wang, Lee, Brennan, and Kolen (2008).

When compared to the unsmoothed equating results, both the log-linear presmoothing and cubic spline postsmoothing methods lead to reduction in variance and mean squared error. Bias tends to be larger for the presmoothed equating methods than the postsmoothed equating methods using $S = .1$, whereas the opposite is true for the variance. However, when a different smoothing parameter is used for the postsmoothed methods (i.e., $S = .3$), the postsmoothed methods tend to outperform the presmoothed methods both in the mean squared error and variance. Given the limited number of simulation conditions used in this study, it seems premature to conclude which smoothing method performs better due to their high dependency on the selected smoothing parameters. A more comprehensive simulation study would be necessary. Using a random groups equating design, Hanson, Zeng, and Colton (1994) found that presmoothing and postsmoothing methods provided comparable levels of performance in terms of equating error.

As summarized in Table 8, the level of correlation between the multiple-choice and free-response constructs and group ability difference that are required to achieve adequate equating by the DTM criterion depends upon equating methods and types of score scales of interest. The frequency estimation methods require a higher correlation and smaller group ability difference than do the chained equipercentile methods. Depending upon the type of primary score scale that is reported to examinees, different decisions would be made on determining an appropriate level of the correlation and group ability difference because the score scales differ in the number of score points, the degree of linearity of the transformation from the raw scores, and the pattern of conditional equating errors. Based on the findings from this simulation, some practical guidelines, somewhat oversimplified, can be developed. Based on the DTM criterion, when the effect size is .2 or smaller the chained equipercentile equating methods seem to provide adequate equating under the following conditions: a) The disattenuated correlation can be as low as .8 for raw-score equating; and b) any level of disattenuated correlation (as low as .5) seems to work for scale-score equating. When the effect size is small (e.g., .05 or lower), all equating methods provide adequate equating even with the very low disattenuated correlation of .5.

It is important to realize that the criterion used in this study to identify an acceptable level of equating is based on the comparison of an arbitrarily selected value of the DTM and bias. A half point of a reported scale-score point has been widely used as a DTM value in many previous studies. The primary reason for using the bias as the criterion in the present study, rather than the mean squared error, is its conceptual similarity to the DTM—namely, both statistics are concerned with “differences.” It is often viewed that the mean squared error is the aggregate error containing both bias and variance (i.e., random sampling error) and thus should be considered as the criterion. In certain cases, however, it would be desirable to consider the two statistics (bias and variance) separately and emphasize one statistic (e.g., bias) over the other (e.g., variance). For example, if the variance is treated as the primary criterion, the frequency estimation methods would be preferable to the chained equipercentile methods in most cases. In this study, all three statistics are considered in evaluating the relative performance of various equating methods, and the bias statistic is used in conjunction with the DTM to determine an acceptable level of equating.

Some limitations of the current study should be recognized. One particular mixed-format test was used to define the numbers of MC, FR, and common items as the basis for simulation. Even though the structure of the test used in this study is typical, care must be exercised when generalizing the results of the present study to specific testing conditions using other mixed-format tests. For example, this study involves a mixed-format test that consists of two different item types, i.e., multiple choice and free response items, and thus the results of this study may not be immediately generalizable to a mixed-format test consisting of more than two item types. Another limitation of this simulation study lies in the assumption that the correlation between the multiple choice and free response sections is the same for the new and old form groups. Also, group ability differences are the same on the two item types in this simulation. It would thus be interesting future research to examine the effects on equating of different multidimensional latent structures for the two populations. Only one sample-size condition is considered in this study. Assuming the same value of DTM is used, the conclusions about the minimally required correlation may well be altered if a larger or smaller sample size is used. Although the sample size of 3,000 could be argued to be “sufficiently large” for many practical equating situations, it would be useful to know how the amount of error changes with different sample sizes. Finally, this study assumes that the multidimensional IRT simple structure model holds in the simulation, which might influence the results of the FE and CE methods that are based on assumptions other than IRT. Despite these limitations, this study offers useful guidance to practitioners.

References

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

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

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.

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

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: American College Testing.

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

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

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). New York: American Council on Education and Macmillan.

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 (Ed.), *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.

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,* 453-461.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating scale data* [computer program]. Chicago, IL: Scientific Software.

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.

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 o Measurement in education, San Diego, CA.

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

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 o Measurement in education, San Diego, CA.

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

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

Table 1

Old-Form Conversion Table

Raw	NSS	Grade
0	0.00	1
1	0.00	1
2	0.00	1
3	0.00	1
4	0.00	1
5	0.00	1
6	0.00	1
7	0.00	1
8	0.00	1
9	0.00	1
10	0.00	1
11	0.00	1
12	0.00	1
13	1.12	1
14	3.62	1
15	5.60	1
16	7.24	1
17	8.49	1
18	9.23	1
19	9.92	1
20	10.84	1
21	11.81	1
22	12.71	1
23	13.60	1
24	14.45	1
25	15.20	1
26	15.89	1
27	16.55	1
28	17.16	1
29	17.74	1
30	18.36	1
31	18.94	1
32	19.49	1
33	20.08	1
34	20.63	1
35	21.18	1
36	21.72	1
37	22.23	1
38	22.76	1
39	23.25	1
40	23.74	1
41	24.26	1
42	24.78	1
43	25.28	1
44	25.74	1
45	26.24	1
46	26.77	1
47	27.24	2
48	27.72	2
49	28.21	2
50	28.68	2
51	29.13	2
52	29.59	2
53	30.06	2
54	30.52	2
55	30.99	2
56	31.47	2
57	31.95	2
58	32.43	2
59	32.90	2
60	33.37	2
61	33.82	3
62	34.28	3
63	34.76	3
64	35.25	3
65	35.73	3
66	36.19	3
67	36.66	3
68	37.15	3
69	37.65	3
70	38.13	3
71	38.59	3
72	39.08	3
73	39.56	3
74	40.06	3
75	40.56	3
76	41.03	4
77	41.53	4
78	42.05	4
79	42.56	4
80	43.06	4
81	43.55	4
82	44.05	4
83	44.53	4
84	45.03	4
85	45.59	4
86	46.13	4
87	46.69	4
88	47.32	5
89	47.88	5
90	48.42	5
91	49.01	5
92	49.55	5
93	50.10	5
94	50.68	5
95	51.25	5
96	51.86	5
97	52.52	5
98	53.15	5
99	53.78	5
100	54.51	5
101	55.27	5
102	56.01	5
103	56.69	5
104	57.35	5
105	58.00	5
106	58.59	5
107	59.26	5
108	60.13	5
109	61.08	5
110	62.03	5
111	62.89	5
112	63.56	5
113	64.68	5
114	66.27	5
115	67.75	5
116	68.87	5
117	69.37	5
118	70.00	5
119	70.00	5
120	70.00	5
121	70.00	5
122	70.00	5
123	70.00	5
124	70.00	5

Note. NSS = normalized scale score

Table 2

Raw Score Summary Statistics for ES = .05

MSE

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.421	0.523	0.296	0.329	0.335	0.400
ρ = 0.95	0.432	0.500	0.311	0.315	0.348	0.376
ρ = 0.90	0.426	0.500	0.286	0.301	0.330	0.370
ρ = 0.85	0.439	0.532	0.316	0.358	0.348	0.405
ρ = 0.80	0.463	0.515	0.334	0.331	0.372	0.391
ρ = 0.75	0.496	0.580	0.386	0.415	0.401	0.449
ρ = 0.70	0.502	0.557	0.391	0.386	0.408	0.424
ρ = 0.65	0.471	0.585	0.352	0.405	0.371	0.447
ρ = 0.60	0.532	0.594	0.417	0.420	0.430	0.453
ρ = 0.55	0.471	0.545	0.365	0.364	0.367	0.400
ρ = 0.50	0.536	0.585	0.411	0.397	0.431	0.436

Squared Bias

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.041	0.018	0.044	0.013	0.038	0.010
ρ = 0.95	0.044	0.014	0.052	0.021	0.042	0.008
ρ = 0.90	0.049	0.017	0.055	0.020	0.046	0.011
ρ = 0.85	0.081	0.033	0.098	0.045	0.078	0.025
ρ = 0.80	0.087	0.031	0.099	0.040	0.084	0.026
ρ = 0.75	0.079	0.032	0.108	0.057	0.076	0.026
ρ = 0.70	0.103	0.034	0.120	0.050	0.100	0.028
ρ = 0.65	0.085	0.029	0.111	0.050	0.081	0.023
ρ = 0.60	0.123	0.049	0.154	0.072	0.119	0.039
ρ = 0.55	0.097	0.030	0.132	0.057	0.094	0.024
ρ = 0.50	0.127	0.040	0.149	0.058	0.123	0.033

Variance

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.380	0.505	0.252	0.316	0.297	0.390
ρ = 0.95	0.388	0.486	0.259	0.294	0.306	0.367
ρ = 0.90	0.377	0.483	0.231	0.280	0.284	0.359
ρ = 0.85	0.357	0.499	0.218	0.313	0.270	0.380
ρ = 0.80	0.376	0.484	0.235	0.291	0.288	0.365
ρ = 0.75	0.417	0.547	0.278	0.358	0.325	0.423
ρ = 0.70	0.399	0.523	0.271	0.336	0.308	0.396
ρ = 0.65	0.386	0.556	0.241	0.355	0.290	0.424
ρ = 0.60	0.409	0.545	0.263	0.348	0.312	0.415
ρ = 0.55	0.374	0.515	0.233	0.307	0.274	0.376
ρ = 0.50	0.408	0.545	0.263	0.339	0.307	0.403

Table 3

Raw Score Summary Statistics for ES = .1

MSE

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.494	0.492	0.364	0.299	0.403	0.367
ρ = 0.95	0.543	0.520	0.413	0.328	0.451	0.392
ρ = 0.90	0.582	0.541	0.452	0.352	0.483	0.408
ρ = 0.85	0.627	0.567	0.501	0.382	0.529	0.434
ρ = 0.80	0.666	0.588	0.544	0.403	0.567	0.452
ρ = 0.75	0.706	0.609	0.592	0.431	0.607	0.471
ρ = 0.70	0.748	0.634	0.636	0.455	0.649	0.494
ρ = 0.65	0.791	0.654	0.682	0.479	0.693	0.512
ρ = 0.60	0.839	0.678	0.736	0.510	0.740	0.534
ρ = 0.55	0.895	0.705	0.793	0.542	0.792	0.559
ρ = 0.50	0.949	0.732	0.852	0.572	0.845	0.583

Squared Bias

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.158	0.043	0.162	0.041	0.153	0.036
ρ = 0.95	0.194	0.057	0.202	0.057	0.191	0.050
ρ = 0.90	0.230	0.070	0.240	0.073	0.225	0.062
ρ = 0.85	0.270	0.084	0.284	0.091	0.265	0.076
ρ = 0.80	0.310	0.098	0.327	0.108	0.305	0.090
ρ = 0.75	0.348	0.111	0.368	0.125	0.343	0.103
ρ = 0.70	0.391	0.127	0.415	0.144	0.386	0.119
ρ = 0.65	0.433	0.142	0.462	0.164	0.429	0.134
ρ = 0.60	0.478	0.159	0.511	0.185	0.473	0.151
ρ = 0.55	0.530	0.180	0.567	0.210	0.524	0.170
ρ = 0.50	0.585	0.204	0.626	0.237	0.579	0.193

Variance

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.336	0.449	0.202	0.258	0.250	0.330
ρ = 0.95	0.349	0.463	0.211	0.271	0.260	0.342
ρ = 0.90	0.352	0.471	0.212	0.279	0.257	0.346
ρ = 0.85	0.358	0.484	0.217	0.291	0.264	0.359
ρ = 0.80	0.356	0.490	0.217	0.295	0.261	0.362
ρ = 0.75	0.358	0.498	0.223	0.306	0.263	0.368
ρ = 0.70	0.358	0.507	0.221	0.311	0.263	0.374
ρ = 0.65	0.358	0.511	0.221	0.316	0.264	0.378
ρ = 0.60	0.361	0.518	0.225	0.325	0.267	0.383
ρ = 0.55	0.364	0.524	0.226	0.332	0.268	0.389
ρ = 0.50	0.364	0.529	0.226	0.335	0.266	0.390

Table 4

Raw Score Summary Statistics for ES = .2

MSE

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	1.002	0.658	0.876	0.439	0.908	0.527
ρ = 0.95	1.032	0.642	0.915	0.429	0.944	0.512
ρ = 0.90	1.158	0.641	1.050	0.465	1.068	0.513
ρ = 0.85	1.294	0.726	1.167	0.532	1.199	0.591
ρ = 0.80	1.467	0.774	1.356	0.589	1.371	0.643
ρ = 0.75	1.717	0.835	1.600	0.652	1.616	0.693
ρ = 0.70	1.843	0.866	1.728	0.688	1.742	0.725
ρ = 0.65	1.882	0.868	1.773	0.704	1.783	0.727
ρ = 0.60	2.213	1.041	2.091	0.848	2.108	0.892
ρ = 0.55	2.369	1.094	2.249	0.910	2.273	0.951
ρ = 0.50	2.353	0.983	2.236	0.807	2.240	0.826

Squared Bias

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.601	0.112	0.626	0.112	0.595	0.104
ρ = 0.95	0.640	0.121	0.659	0.122	0.635	0.113
ρ = 0.90	0.785	0.168	0.814	0.178	0.782	0.159
ρ = 0.85	0.884	0.176	0.909	0.190	0.879	0.169
ρ = 0.80	1.073	0.247	1.097	0.249	1.068	0.238
ρ = 0.75	1.334	0.330	1.358	0.341	1.328	0.320
ρ = 0.70	1.465	0.377	1.490	0.393	1.462	0.370
ρ = 0.65	1.477	0.329	1.498	0.348	1.473	0.321
ρ = 0.60	1.810	0.496	1.834	0.508	1.801	0.484
ρ = 0.55	1.972	0.538	1.994	0.552	1.968	0.529
ρ = 0.50	1.980	0.476	1.999	0.494	1.971	0.464

Variance

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.401	0.546	0.250	0.327	0.313	0.424
ρ = 0.95	0.391	0.521	0.255	0.306	0.308	0.399
ρ = 0.90	0.373	0.473	0.236	0.287	0.285	0.354
ρ = 0.85	0.410	0.550	0.258	0.342	0.321	0.422
ρ = 0.80	0.394	0.527	0.259	0.340	0.303	0.405
ρ = 0.75	0.384	0.505	0.241	0.311	0.288	0.373
ρ = 0.70	0.378	0.489	0.238	0.295	0.280	0.356
ρ = 0.65	0.404	0.540	0.275	0.356	0.310	0.406
ρ = 0.60	0.403	0.544	0.258	0.339	0.307	0.408
ρ = 0.55	0.398	0.556	0.255	0.359	0.306	0.422
ρ = 0.50	0.373	0.508	0.236	0.314	0.269	0.363

Table 5

Raw Score Summary Statistics for ES = .3

MSE

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	1.642	0.698	1.550	0.500	1.549	0.564
ρ = 0.95	1.942	0.788	1.830	0.584	1.837	0.652
ρ = 0.90	2.238	0.881	2.128	0.679	2.133	0.738
ρ = 0.85	2.557	0.978	2.444	0.781	2.451	0.836
ρ = 0.80	2.894	1.082	2.784	0.893	2.789	0.939
ρ = 0.75	3.270	1.217	3.156	1.027	3.164	1.066
ρ = 0.70	3.646	1.345	3.527	1.159	3.538	1.193
ρ = 0.65	4.043	1.481	3.920	1.300	3.934	1.326
ρ = 0.60	4.480	1.643	4.353	1.464	4.368	1.485
ρ = 0.55	4.903	1.792	4.771	1.617	4.791	1.632
ρ = 0.50	5.359	1.960	5.224	1.787	5.247	1.799

Squared Bias

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	1.293	0.239	1.338	0.240	1.285	0.228
ρ = 0.95	1.580	0.317	1.619	0.320	1.570	0.306
ρ = 0.90	1.875	0.401	1.913	0.407	1.867	0.389
ρ = 0.85	2.193	0.493	2.227	0.502	2.183	0.480
ρ = 0.80	2.535	0.597	2.565	0.609	2.524	0.583
ρ = 0.75	2.901	0.712	2.925	0.725	2.891	0.697
ρ = 0.70	3.277	0.831	3.297	0.846	3.265	0.815
ρ = 0.65	3.672	0.961	3.688	0.978	3.658	0.941
ρ = 0.60	4.103	1.109	4.115	1.130	4.087	1.089
ρ = 0.55	4.525	1.253	4.532	1.277	4.506	1.230
ρ = 0.50	4.981	1.418	4.987	1.445	4.965	1.396

Variance

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
ρ = 1.00	0.349	0.459	0.211	0.260	0.264	0.336
ρ = 0.95	0.362	0.471	0.211	0.263	0.267	0.345
ρ = 0.90	0.363	0.480	0.216	0.272	0.266	0.349
ρ = 0.85	0.363	0.486	0.217	0.279	0.269	0.356
ρ = 0.80	0.359	0.485	0.218	0.284	0.265	0.356
ρ = 0.75	0.369	0.505	0.231	0.302	0.273	0.370
ρ = 0.70	0.369	0.513	0.230	0.312	0.274	0.378
ρ = 0.65	0.371	0.520	0.232	0.322	0.276	0.385
ρ = 0.60	0.377	0.533	0.238	0.334	0.281	0.396
ρ = 0.55	0.379	0.539	0.239	0.340	0.285	0.403
ρ = 0.50	0.378	0.542	0.238	0.342	0.282	0.404

Table 6

Raw Score Summary Statistics for ES = .5

MSE

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
$\rho = 1.00$	3.974	1.151	4.004	0.933	3.876	1.006
$\rho = 0.95$	4.537	1.240	4.533	1.044	4.441	1.102
$\rho = 0.90$	5.516	1.577	5.544	1.357	5.411	1.419
$\rho = 0.85$	6.372	1.800	6.378	1.583	6.267	1.639
$\rho = 0.80$	7.049	1.916	6.981	1.710	6.935	1.758
$\rho = 0.75$	8.286	2.268	8.215	2.084	8.182	2.104
$\rho = 0.70$	9.412	2.658	9.317	2.478	9.305	2.500
$\rho = 0.65$	10.246	2.918	10.105	2.713	10.133	2.752
$\rho = 0.60$	11.656	3.399	11.489	3.186	11.531	3.212
$\rho = 0.55$	13.099	3.896	12.889	3.667	12.970	3.714
$\rho = 0.50$	14.163	4.273	13.944	4.054	14.043	4.086

Squared Bias

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
$\rho = 1.00$	3.574	0.621	3.762	0.627	3.564	0.598
$\rho = 0.95$	4.125	0.701	4.274	0.720	4.117	0.685
$\rho = 0.90$	5.115	0.988	5.288	0.993	5.102	0.966
$\rho = 0.85$	5.975	1.210	6.123	1.211	5.959	1.186
$\rho = 0.80$	6.606	1.321	6.699	1.330	6.588	1.300
$\rho = 0.75$	7.886	1.725	7.960	1.750	7.877	1.700
$\rho = 0.70$	8.970	2.082	9.024	2.096	8.955	2.057
$\rho = 0.65$	9.832	2.340	9.850	2.350	9.815	2.311
$\rho = 0.60$	11.219	2.785	11.200	2.786	11.192	2.748
$\rho = 0.55$	12.654	3.283	12.605	3.284	12.629	3.254
$\rho = 0.50$	13.706	3.656	13.646	3.650	13.687	3.619

Variance

	Equating Methods					
Correlation	UnSm_FE	UnSm_CE	PreSm_FE	PreSm_CE	PostSm_FE	PostSm_CE
$\rho = 1.00$	0.399	0.530	0.242	0.306	0.312	0.408
$\rho = 0.95$	0.412	0.539	0.259	0.324	0.324	0.417
$\rho = 0.90$	0.401	0.589	0.255	0.364	0.309	0.453
$\rho = 0.85$	0.397	0.590	0.255	0.372	0.308	0.453
$\rho = 0.80$	0.444	0.596	0.282	0.379	0.348	0.459
$\rho = 0.75$	0.400	0.543	0.255	0.334	0.305	0.404
$\rho = 0.70$	0.442	0.576	0.293	0.382	0.350	0.443
$\rho = 0.65$	0.414	0.578	0.256	0.364	0.318	0.441
$\rho = 0.60$	0.437	0.614	0.289	0.400	0.340	0.464
$\rho = 0.55$	0.445	0.612	0.284	0.384	0.340	0.461
$\rho = 0.50$	0.457	0.617	0.298	0.404	0.356	0.467

Table 7

Raw Score Summary Statistics Using S = .3 for Postsmoothing

	FE (*S*=.1)	CE (*S*=.1)	PreSm_FE	PreSm_CE	FE (*S*=.3)	CE (*S*=.3)
			ES = .05			
MSE						
ρ = 1.00	0.335	0.400	0.296	0.329	0.267	0.302
ρ = 0.80	0.372	0.391	0.334	0.331	0.297	0.290
ρ = 0.50	0.431	0.436	0.411	0.397	0.364	0.339
Squared Bias						
ρ = 1.00	0.038	0.010	0.044	0.013	0.040	0.013
ρ = 0.80	0.084	0.026	0.099	0.040	0.083	0.027
ρ = 0.50	0.123	0.033	0.149	0.058	0.129	0.041
Variance						
ρ = 1.00	0.297	0.390	0.252	0.316	0.226	0.290
ρ = 0.80	0.288	0.365	0.235	0.291	0.214	0.263
ρ = 0.50	0.307	0.403	0.263	0.339	0.235	0.298
			ES = .2			
MSE						
ρ = 1.00	0.908	0.527	0.876	0.439	0.827	0.415
ρ = 0.80	1.371	0.643	1.356	0.589	1.286	0.538
ρ = 0.50	2.240	0.826	2.236	0.807	2.159	0.732
Squared Bias						
ρ = 1.00	0.595	0.104	0.626	0.112	0.597	0.109
ρ = 0.80	1.068	0.238	1.097	0.249	1.059	0.242
ρ = 0.50	1.971	0.464	1.999	0.494	1.956	0.468
Variance						
ρ = 1.00	0.313	0.424	0.250	0.327	0.230	0.306
ρ = 0.80	0.303	0.405	0.259	0.340	0.228	0.296
ρ = 0.50	0.269	0.363	0.236	0.314	0.203	0.265
			ES = .5			
MSE						
ρ = 1.00	3.876	1.006	4.004	0.933	3.804	0.903
ρ = 0.80	6.935	1.758	6.981	1.710	6.834	1.657
ρ = 0.50	14.043	4.086	13.944	4.054	13.908	3.981
Squared Bias						
ρ = 1.00	3.564	0.598	3.762	0.627	3.571	0.606
ρ = 0.80	6.588	1.300	6.699	1.330	6.572	1.313
ρ = 0.50	13.687	3.619	13.646	3.650	13.635	3.634
Variance						
ρ = 1.00	0.312	0.408	0.242	0.306	0.233	0.296
ρ = 0.80	0.348	0.459	0.282	0.379	0.263	0.344
ρ = 0.50	0.356	0.467	0.298	0.404	0.274	0.346

Table 8

Minimum Level of Correlation between MC and FR Constructs for Adequate Equating Based on the DTM criterion

Raw Scores

	Effect Size				
Equating Methods	.05	.1	.2	.3	.5
FE	A	.9	N	N	N
CE	A	.7	.8	1.0	N

Normalized Scale Scores

	Effect Size				
Equating Methods	.05	.1	.2	.3	.5
FE	A	A	.85	N	N
CE	A	A	A	.7	.9

AP Grades

	Effect Size				
Equating Methods	.05	.1	.2	.3	.5
FE	A	.7	N	N	N
CE	A	A	A	.85	N

Note. “A” means that the results are acceptable across all correlation levels with the lowest being .5; “N” represents that the methods do not work even with a perfect correlation; FE includes the unsmoothed, presmoothed, and postsmoothed frequency estimation equating methods; and CE includes the unsmoothed, presmoothed, and postsmoothed chained equipercentile equating methods.

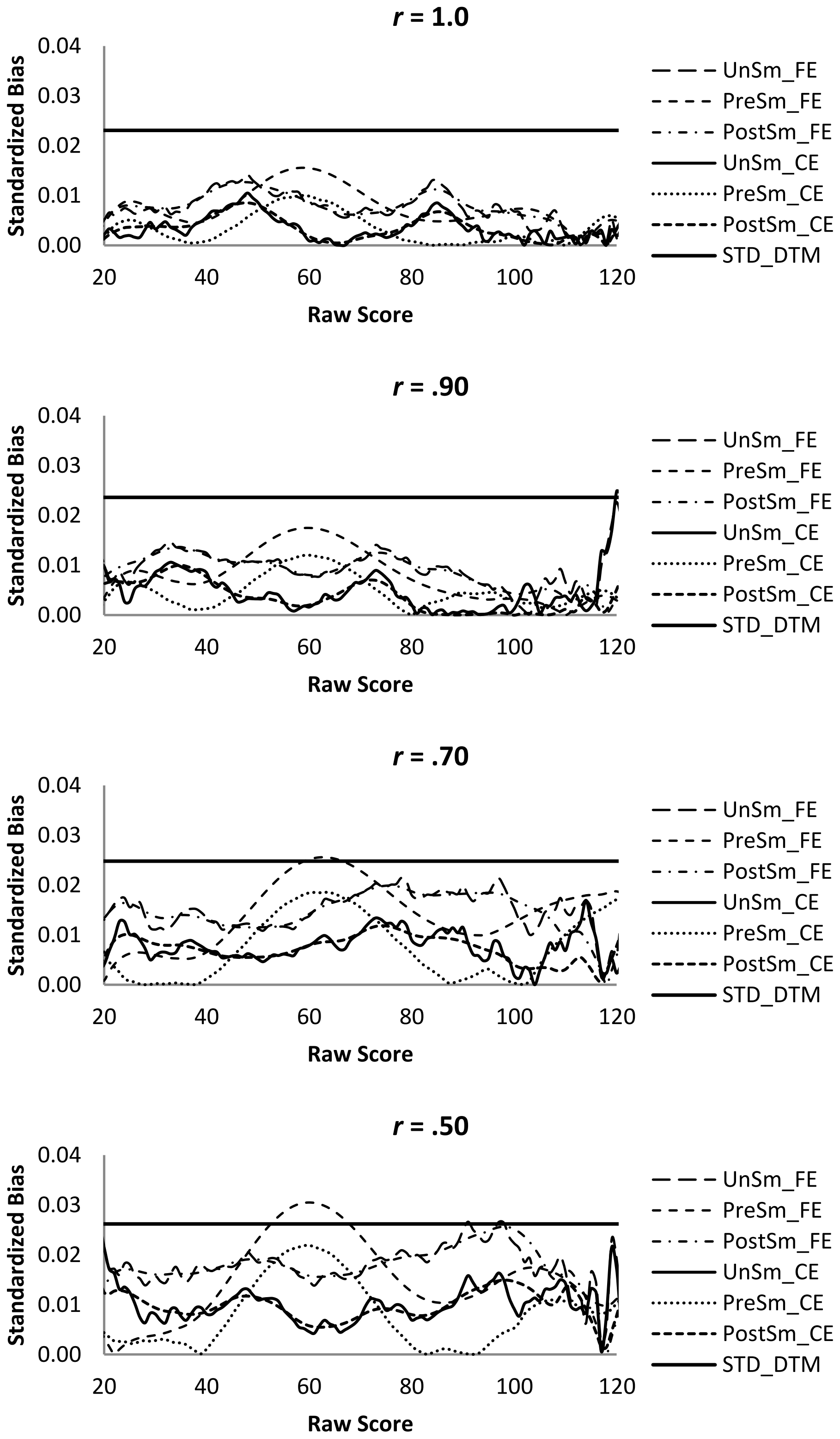


Figure 1. Raw score standardized bias for *ES* = .05.

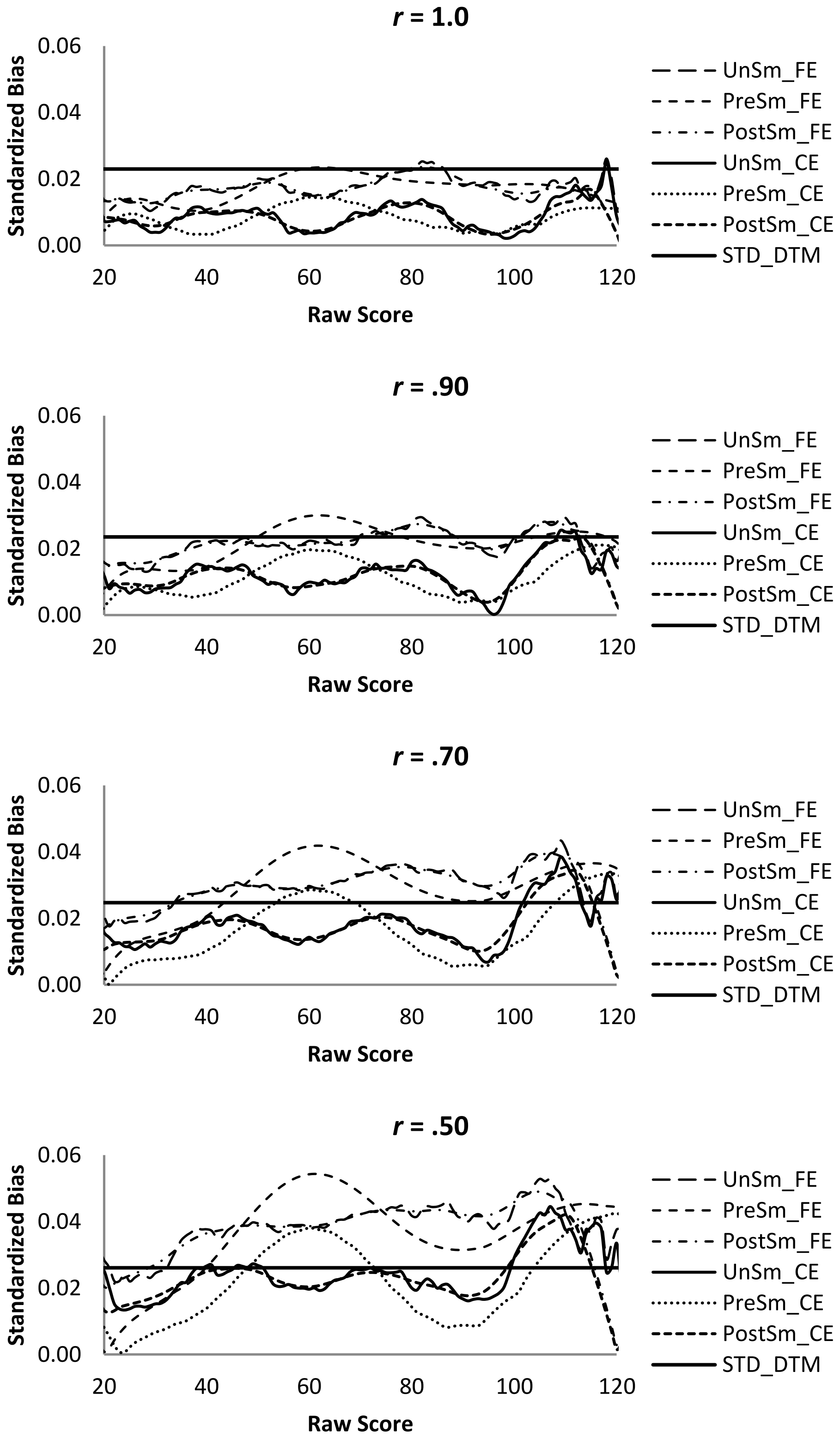


Figure 2. Raw score standardized bias for *ES* = .1.

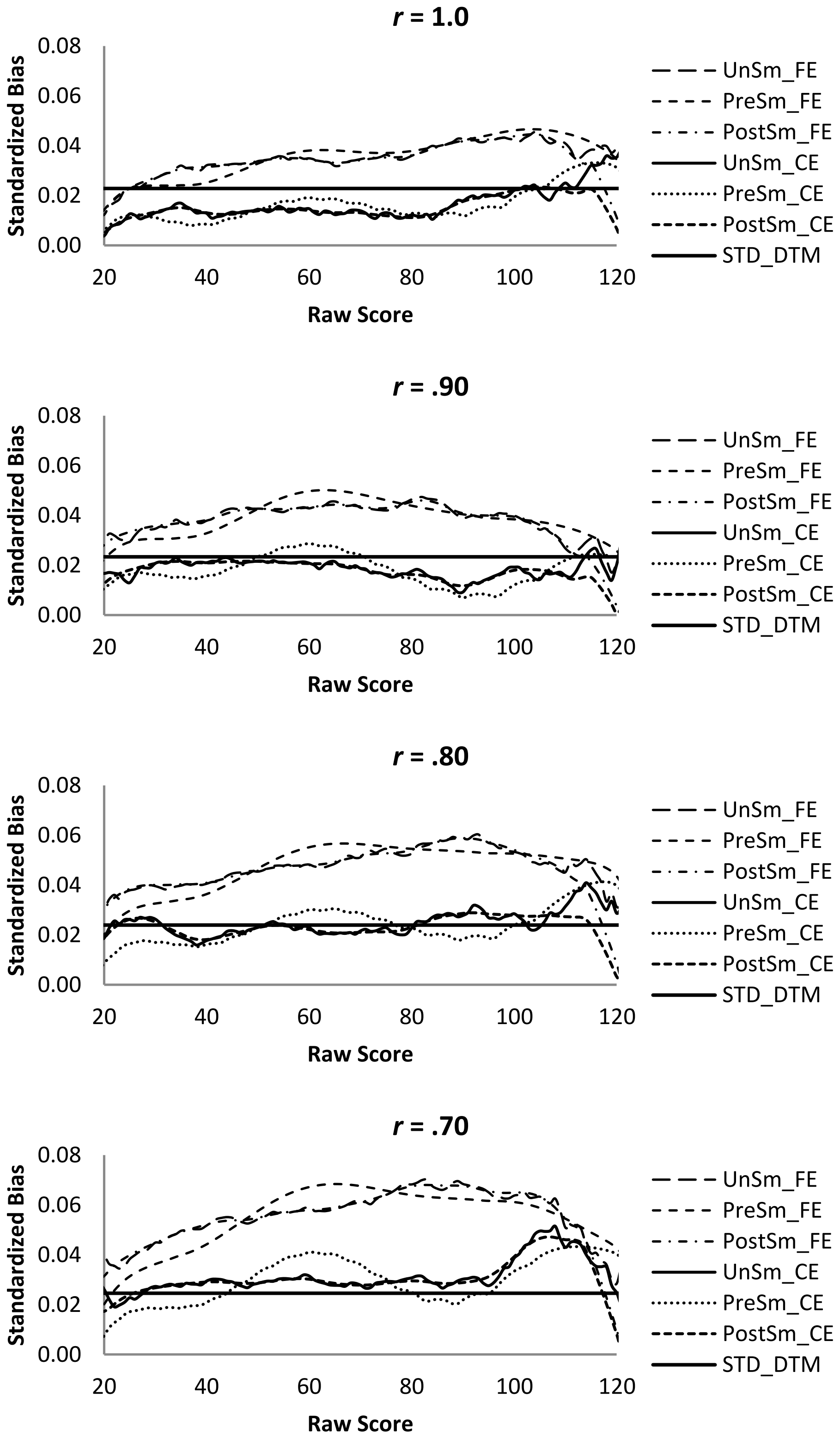


Figure 3. Raw score standardized bias for *ES* = .2.

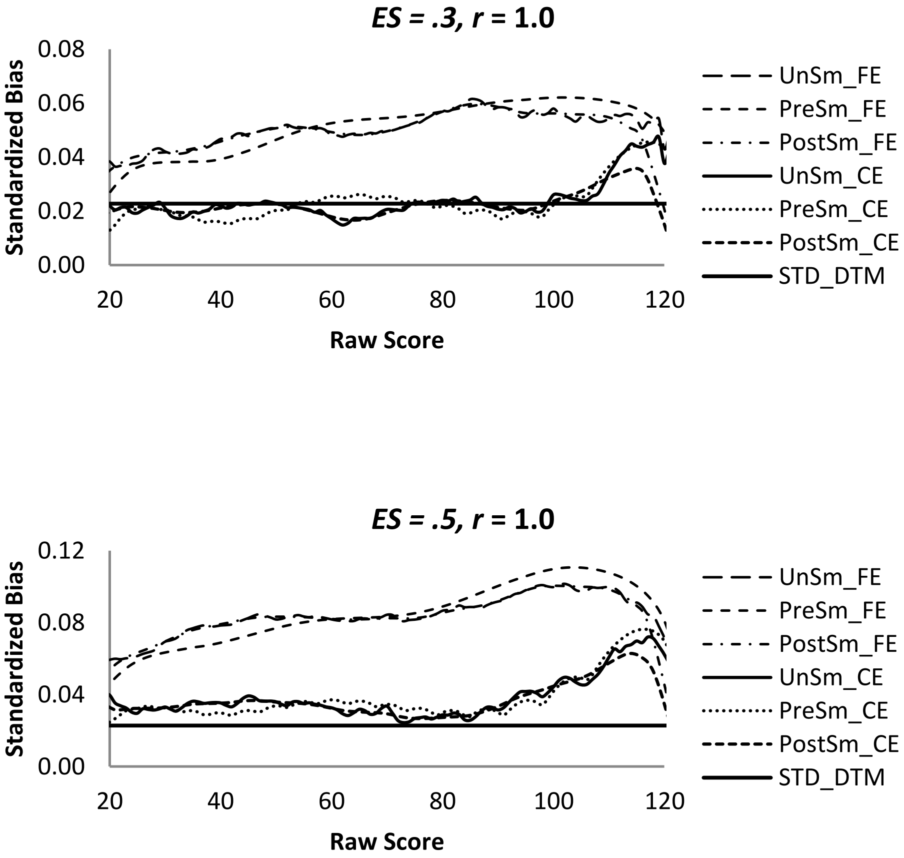


Figure 4. Raw score standardized bias for *ES* = .3 and *ES* = .5 under perfect correlation.

Chapter 3: Effects of Group Differences on Equating Using Operational and Pseudo-Tests

Sarah L. Hagge and Michael J. Kolen

The University of Iowa, Iowa City, IA

Abstract

The primary goals of this study were to examine how group differences impact accuracy of equating results and to understand whether analyses conducted on operational test forms and pseudo-test forms lead to the same results and conclusions. This study examined tests in three different subject areas (English Language, Spanish Language, and Chemistry) that were presumed to represent different degrees of dimensionality. Proficiency difference between old and new form examinee groups and equating method were also varied. The results of the study confirm those of previous studies that equating tends to be less biased when old and new form examinee groups are similar in proficiency. In addition, chained equipercentile and IRT equating methods were found to be less biased than frequency estimation when proficiency differences were large. However, standard errors of equating tended to be larger for chained equipercentile equating. Results were similar across operational and pseudo-test form criteria.

Effects of Group Differences on Equating using Operational and Pseudo-Tests

The use of mixed-format tests containing both multiple-choice (MC) and free-response (FR) items has increased in recent years. However, much of the research on equating has been conducted on MC-only tests. The first goal of this study was to gain increased understanding about the extent to which characteristics of operational mixed-format tests affect equating results. A second purpose of this study was to understand whether similar analyses on two different classes of data lead to the same results and conclusions. Specifically, this study addresses the following questions as they pertain to mixed-format tests:

1. What is the impact on equated scores when examinees on one mixed-format test form are higher in proficiency, as measured by the items in common between test forms, than examinees on the other mixed-format test form?
2. How much do equated scores vary across equating methods?
3. How do the content and statistical specifications of a test (e.g., subject area, correlation between MC and FR scores) impact equated scores?
4. To what extent do analyses with two different classes of data, operational test forms and pseudo-test forms, result in the same findings?

Theoretical Framework

MC items are a type of selected response item with a question or statement stem followed by possible answer choices (Ferrara & DeMauro, 2006). Many of the strengths of MC items lie in efficiency of administration and scoring. However, MC items have been criticized for their perceived inability to assess higher-order thinking skills and curricular overemphasis on MC specific test-taking strategies. FR items require examinees to produce an answer or product without existing answer choices. Some of the strengths of FR items include that they are easy to create relative to MC items and place a curricular emphasis on writing and away from memorization and recall (Clauser, Margolis, & Case, 2006; Ebel & Frisbie, 1991; Ferrara & DeMauro, 2006). However, FR items are more expensive to score, easier to memorize, and scores on tests consisting of CR items typically have lower reliability. Fewer tasks can be administered during an administration period, resulting in limited sampling of the content domain (Ebel & Frisbie, 1991; Ferrara & DeMauro, 2006).

A balance between MC and FR items appears to have been found with mixed-format tests that contain both MC and FR items. By including each item format on a test, some of the weaknesses of each format are mitigated by strengths of the other. However, the combination of both item formats introduces potential problems for existing psychometric methodologies.

Definition of Equating

"*Equating* is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content" (Kolen & Brennan, 2004, p. 2). A variety of data collection designs and methodologies for equating have been developed. Three data collection designs are commonly used: single group design, random groups design, and the common-item nonequivalent groups (CINEG) design. The focus of the current study is on the CINEG design. In the CINEG design, some items are selected to be administered on two test forms. These items are referred to as common items. Two groups of examinees that are not assumed to be equivalent in overall proficiency take one of two test forms. The common items allow for separation of differences in examinee score distributions across the two test forms into form difficulty and examinee proficiency. The CINEG design is very flexible and widely used, but it requires strong statistical assumptions and careful development of a set of common items.

Equating Research Data Classes

Studies on equating have used various classes of data for investigating equating relationships, including operational, pseudo, and simulated test forms. It is of interest to know whether the different classes of data result in the same conclusions. The primary benefit of using operational test forms is that the data consist of items, test forms, and examinees from an actual test administration. However, there is no clear criterion for evaluating the adequacy of equating. Equating methods can be compared, but there is no way of knowing which method is more accurate than another. In contrast, with simulated test forms, the population is known, because all of the items and examinee responses have been generated. Examinee and item characteristics can be manipulated in order to create tests that align with the problems the researcher is trying to solve. Although simulated test forms are typically based on operational test forms, concerns exist about the extent to which simulated test forms reflect operational test forms.

A third class of data is pseudo-test forms (Holland, Sinharay, von Davier, & Han, 2008; Ricker & von Davier, 2007; Sinharay & Holland, 2007). Pseudo-test forms are created by splitting the items from one test form from an operational administration in half to create two test forms. A pseudo-test form uses an operational test form and creates a reasonable criterion, because data exist for the same examinees on two pseudo-test forms. Additionally, pseudo-test forms allow the researcher to manipulate the composition of items on the test forms. However, pseudo-test forms may not accurately represent operational test forms for a number of reasons. Pseudo-test forms are shortened forms of the original test. Further, pseudo-test forms are intended to be parallel in content and statistical specifications, but it is plausible that this parallelism may not be achieved in practice. Also, pseudo-test forms are typically created in a way to address a particular problem, which may or may not reflect the way operational test forms exist in practice.

Overview of Relevant Research

Although much of the research has been conducted on tests containing only MC items, a growing body of literature has focused on FR-only tests and mixed-format tests. A number of consistent results have been found across the equating studies that have been conducted.

Test and examinee characteristics. One factor that may impact equating is the extent to which MC and FR items are measuring equivalent constructs on a given test, which may differ according to the subject area as well as format of the FR item (Traub, 1993). Research also suggests that when MC and FR items are designed to measure equivalent constructs, they perform similarly (Rodriguez, 2003).

Test characteristics also impact equating results, although results may be dependent on differences in examinee proficiency across test forms. Sinharay and Holland (2007) investigated the impact of statistical representativeness of common-item sets on equating using simulated data. They found the effect of common-item set had a much smaller impact on equating than equating method, sample size, group differences, or test length. Wu, Huang, Huh, and Harris (2009) examined the effectiveness of using MC items as an external common-item set for a FR test with simulated data. They found that when the correlation between MC and FR items was lower, greater bias occurred in the equating relationship. Additionally, as the mean difference between groups increased, the amount of bias in the sample equating relationship also increased. Cao (2008) simulated a number of characteristics of a mixed-format test. Bias was always smaller and classification consistency was always higher for the equivalent groups condition.

Equating methods and designs. For MC tests, frequency estimation (FE) and chained equipercentile (CE) equating methods typically performed similarly when examinee groups were similar in proficiency (Wang, Lee, Brennan, & Kolen, 2008), but led to different results when examinee groups differed in proficiency (Harris & Kolen, 1990). Specifically, when groups differed substantially, CE produced somewhat more accurate equating results than FE (Holland, et al., 2008; Sinharay & Holland, 2007; Wang, et al., 2008). Similar results were found for a mixed-format test study (Lee, He, Hagge, Wang, & Kolen, 2012). In research for both MC and mixed-format tests, FE has been found to result in smaller standard errors than CE (Lee, et al., 2012; Sinharay & Holland, 2007; Wang, et al., 2008).

Although many of the comparisons of IRT and traditional equating methods have been conducted for random group designs, there is evidence to suggest IRT equating results might lead to more stable or accurate equating results than equipercentile methods (Han, Kolen, & Pohlmann, 1997). However, there is also evidence to suggest that IRT and traditional equating methods lead to similar results (von Davier & Wilson, 2008; Harris & Kolen, 1986). One study also found that standard errors of equating were smaller for IRT observed score (OS) equating as compared to IRT true score (TS) equating (Tsai, Hanson, Kolen, & Forsyth, 2001).

Equating research data classes. As discussed previously, data classes commonly include one of three approaches: operational test forms, pseudo-test forms, or simulated test forms. For operational test forms, it is common in equating studies to conduct various equatings and compare results across the equating methods (Harris & Crouse, 1993). For this method, there is generally no criterion to judge the source of the different results across equating methods. Pseudo-test forms appear to be a relatively recent method of assessing equating results (Holland, et al., 2008; Ricker & von Davier, 2007; Sinharay & Holland, 2007). Commonly, a single-group equating relationship is used as the criterion relationship for pseudo-test forms, because data exist on both pseudo-test forms for all examinees. For simulated test forms, equating results are typically evaluated by how well the true population relationship is recovered (Harris & Crouse, 1993). However, if the model used to generate the data is also implemented as a study condition, results may be biased favorably towards the generating model.

The extent to which findings and conclusions varied as a result of the class of data used in the study was difficult to determine from the literature reviewed. The primary classes of data in the studies reviewed were simulated test forms and pseudo-test forms, and for the most part, studies resulted in similar conclusions. However, because the majority of the studies did not investigate multiple classes of data within a study, it is impossible to tell whether differences in findings were the result of the class of data or the study characteristics. Sinharay and Holland

(2006, 2007) were one exception. They used multiple classes of data in their studies and found similar results. However, their research was conducted on MC-only tests.

Although a great deal of research has been conducted on equating, much of the research has focused on MC-only tests. Additionally, many of the studies have focused primarily on simulated test forms or on one pseudo-test subject area. Typically, traditional and IRT equating methods have been investigated separately. The purpose of this study was to contribute to current literature on equating mixed-format tests. Specifically, this study examined tests in three different subject areas that were presumed to represent different degrees of dimensionality. Further, these tests were comprised of different numbers and types of MC and FR items. Both traditional and IRT equating methods were considered. Lastly, this study incorporated both operational and pseudo-test forms to evaluate how findings may differ across classes of data.

Methodology

The methodology describes the original operational test forms used for the study, data preparation, dimensionality assessment, and factors of investigation.

Selection of Tests

Data for this study were from College Board Advanced Placement (AP) tests. It is important to note that, although Advanced Placement (AP) tests were used for analyses in this study, the exams were manipulated in order to investigate how equating methods, test characteristics, and differences in examinee group proficiency affect equating results for mixed-format tests. The tests were modified in such a way that the characteristics of the tests and groups of examinees no longer represented the AP tests as administered operationally. Consequently, generalizations of the results and findings from this study should not be made to AP tests.

Subject area, MC and FR correlations, and number of examinees were considered when selecting the specific AP tests for this study. A variety of subject area tests were desired in order to determine whether similar patterns of findings occurred across different subject areas. Three tests were selected spanning science and language subject areas: English Language, Spanish Language, and Chemistry. For this study, one operational equating relationship and corresponding datasets were selected for each of the three tests.

All three of the tests were mixed-format tests, meaning the tests contained MC items as well as at least one type of FR item format. The English Language tests were comprised of MC items and three longer essay items. The Spanish Language tests addressed listening, reading, writing, and speaking skills. The MC items measured listening and reading comprehension. FR prompts included paragraph completion and word fill-ins, written interpersonal communication and integrated essays, and speaking prompts based on picture sequences and directed responses. The Chemistry tests consisted of MC items covering broad Chemistry topics. The FR items included quantitative and non-quantitative prompts, prompts on writing balanced chemical equations, and an item about reactants. Additionally, examinees were allowed a choice of one of two prompts for two of the FR items for Chemistry 2005.

Table 1 contains a summary of the number of MC and FR items for each test by year of administration. The Spanish Language and Chemistry tests contain multiple FR item formats worth different maximum point values. The first column in Table 1 lists the subject area test, and the second column provides the year of administration. The third column provides the number of MC items on a given form. The fourth column contains the number of items for a given FR item format on a given test form, and the fifth column contains the maximum point values for each FR item format. The last column contains the total number of points on the test forms. For example, for Spanish Language, column four contains the numbers 20, 5, and 2, meaning that there are three item formats. The first format contains 20 items, the second contains five, and the third contains two. Column five contains the numbers 1, 4, and 9. These numbers indicate that the first FR item format is worth a maximum of one point, the second item format is worth a maximum of four points, and the third item format is worth a maximum of nine points. There are 20 items worth one point each, five items worth four points each, and two items worth nine points each. For Chemistry, the 14 point FR item was originally worth 15 points, but one category was collapsed for IRT analyses.

The second factor considered in selecting tests was observed and disattenuated correlations between MC and FR scores. A range of correlations was desired in order to investigate how the MC and FR correlation impacts equated scores. Observed MC and FR correlations were calculated using Pearson correlations and disattenuated MC and FR correlations were calculated using the observed Pearson correlations and coefficient α estimates of reliability. English Language had the lowest observed and disattenutated correlations, followed by Spanish Language. Chemistry was the test selected with the highest correlations. It is important to note that because a different subject area was selected for each level of correlation, subject area and MC and FR correlation were completely confounded. However, the operational AP science exams tended to have higher MC and FR correlations, and the foreign language exams tended to have moderate correlations. These correlations are provided in the results section of this chapter.

Data Preparation

For this study, number-correct scoring was used. However, formula scoring was used for the original operational test forms. Examinees were advised that incorrect answers would be penalized. Consequently, some examinees had missing responses for a large number of MC items. In order to use number-correct scoring, examinees completing fewer than 80% of the MC items were removed. Then, formula scores were transformed to number-correct scores. Incorrect responses were coded as 0, correct responses were coded as 1, and imputation was used for the missing responses. The two-way imputation procedure described by Sijtsma and van der Ark (2003) was used in this study. Descriptive statistics for the imputed data are provided in Table 2. In this table, CI refers to common items and CO to the composite over MC and CR items.

Operationally, weights used for AP exams are complex. Weights differ by item format in order to ensure that each item format is given the intended proportion of points according to test specifications. Additionally, because test forms do not contain the same number of items across test forms, weights ensure the number of score points is the same across years. In this study, to simplify the computation of total scores, summed scores were used. That is, each MC or FR point was worth one point.

Dimensionality Assessment

To evaluate whether the IRT assumption of unidimensionality held, a dimensionality assessment was conducted. For this study, disattenuated correlations between MC and FR scores were considered, and a principal components analysis (PCA) was conducted using tetrachoric and polychoric correlations among all individual items. Disattenuated MC and FR correlations near 1.00 indicated the MC and FR sections of the test were measuring essentially the same dimensions. For the PCA, four guidelines were considered. The first guideline was to retain components with eigenvalues greater than one (Orlando, 2004; Rencher, 2002). A second guideline was to examine the scree plots of eigenvalues for a break between large and small eigenvalues (Orlando, 2004; Rencher, 2002). A third guideline was to compare the ratio of the first and second eigenvalue to the ratio of the second and third eigenvalues for ratios larger than three (Divgi, 1980; Hattie, 1985; Jiao, 2004; Lord, 1980). Lastly, Reckase (1979) recommended that 20% or more of the total variance should be explained by the first principal component in order for a test to be considered unidimensional.

Equating Research Data Class

Operational test forms. The first class of data considered in this study was based on operational test forms. Operational test form analyses were conducted with the test forms intact. However, as described later, different samples of examinees were selected for the operational test form analyses. The procedures described for the operational test form analyses were repeated for each of the three tests selected for the study.

Pseudo-test forms. The three tests used for operational test form analyses were also used for the pseudo-test form analyses, and similar analyses were conducted. Only one year of administration was selected for each of the three tests. Pseudo-test forms were created by splitting one test form into two new test forms. Old and new pseudo-test forms were created from the original operational test forms. Pseudo-test forms were created to be as similar as possible in terms of content and statistical specifications based on the sample of examinees completing at least 80% of the items. Difficulty was calculated for each item as the mean item score over all examinees. For polytomous items, the mean item score was divided by the total number of points possible for the item to put the difficulty on the same scale as that of the MC items. Pearson correlations between each MC or FR item and the total scores were calculated as discrimination. Composition of the pseudo-test forms is included in Table 3, and descriptive statistics are included in Table 4. The pseudo-test forms were created with both MC and FR items in the common-item set. Full CI refers to all items in common, and NFR refers to MC items only (**N**o **FR**). The MC-only common items were used for all analyses in this study. However, sampling, which will be described in the next section, was conducted based on Full CI.

Factors of Investigation

Examinee common-item effect sizes. In order to investigate research question one, it was necessary to create differences in examinee score means across test forms for common-item scores. Differences in common-item score means were created using a sampling process based on demographic variables in order to create various effect sizes across old and new test forms for common-item scores. It was determined that examinees belonging to different demographic groups performed differently across common-items, FR items, and MC items. Therefore, by oversampling certain demographic groups, different common-item effect sizes could be obtained across old and new forms.

Common-item effect sizes (CI ES) indicated the standardized mean common-item scores of the examinees differed across old and new test forms. Three levels of CI ES were chosen for the operational test forms: 0.00, 0.20, and 0.40. The same three levels, plus an additional level of 0.60, were chosen for the pseudo-test forms. As a frame of reference, Kolen and Brennan (2004) indicated that differences of 0.30 standard deviation units or more could result in large differences across equating methods. A single sample of examinees was selected for each of the ES levels. The sampling process was replicated at each ES level. ES were calculated for new form minus old form common-item scores, as shown in Equation 1,

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

where

$\bar{x}_1$ = mean for new form;

$\bar{x}_2$ = mean for old form;

n_1 = number of examinees for new form;

n_2 = number of examinees for old form;

s_1^2 = variance of scores for new form; and

s_2^2 = variance of scores for old form.

Equating methods. Equating was conducted for the CINEG design using four equating methods: frequency estimation (FE), chained equipercentile (CE), IRT true-score (TS), and IRT observed-score (OS). For the traditional equating methods, FE and CE, equating was conducted using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). Cubic spline postsmoothing was conducted, and an *S* value of 0.1 was chosen based on previous research conducted on AP exams. For FE, the new form population received a weight of one to create synthetic populations. The equating process was replicated using 500 replication samples.

Before conducting equating for the IRT methods, item parameter estimates and estimates of the distribution of proficiency were obtained using PARSCALE (Muraki & Bock, 2001). The three-parameter logistic (3PL) model was used for the MC items, and the generalized partial credit model (GPCM) was used to estimate parameters for the FR items. The new form item and proficiency parameter estimates were placed onto the scale of the old form using STUIRT (Kim & Kolen, 2004) and the Haebara method (Haebara, 1980) of scale linking. TS and OS equating were conducted for raw scores using POLYEQUATE (Kolen, 2003). In order to maintain consistency with FE, the new form population received a weight of one for OS. The entire IRT equating process was also replicated using 500 replication samples.

Evaluation

For the operational test forms, a separate criterion equating was established for each of the four equating methods. This criterion equating was for CI 0.00. This ES was selected as the criterion because group differences were closest to zero; consequently, assumptions were least likely to be violated. Additionally, this pattern essentially represented no difference in proficiency between old and new forms. Although this criterion is reasonable, it is not an absolute criterion. It is important to note that the criterion equating relationship differed for each equating method. That is, the criterion equating relationship for FE was calculated using the FE equating method for the CI 0.00 sampling condition. The criterion equating relationship for CE was calculated using the CE equating method for the CI 0.00 sampling condition, and so on. The similarity of these criterion equatings to one another are considered in the results section.

Two pseudo-test forms were created using a single test form; consequently, the same examinees took both pseudo-test forms. Therefore, the criterion equating relationship was established for each subject area test for the pseudo-tests using the single group equating design. After the pseudo-test forms were created, single group equating was conducted prior to sampling subgroups of examinees. A single group equipercentile equating relationship was used as the criterion for FE and CE. TS and OS were used as the criterion equating relationships for TS and OS, respectively. Concurrent calibration was used for estimating item and proficiency parameters for the single group. The similarity of these criterion equatings to one another are considered in the results section.

To evaluate the results from the operational and pseudo-test form analyses bias, root mean squared error (RMSE), and conditional standard error of equating (CSE) were calculated for each score point. Equations 2, 3, and 4 represent these statistics, as follows:

$$Bias_i = \frac{\sum_{j=1}^{J}[\hat{e}_j(x_i) - e^*(x_i)]}{J}, \qquad (2)$$

$$CSE_i = \sqrt{Var_i\left[\hat{e}_j(x_i) - \bar{\hat{e}}(x_i)\right]}\,, \tag{3}$$

and

$$RMSE_i = \sqrt{Bias_i^2 + CSE_i^2}\ . \tag{4}$$

In these three equations, *i* is a score point, *j* is a replication, *J* is the total number of replications, $e^*(x_i)$ is the old form equivalent of a new form raw score for the criterion equating relationship, and $\hat{e}_j(x_i)$ is the old form equivalent of a new form raw score for a study condition equating relationship. 500 replications of the criterion equating relationship were conducted in order to compare the magnitude of the standard errors for the criterion to the standard errors for the study condition equatings. The study conditions were evaluated against the mean criterion equating over 500 replications.

Weighted average root mean squared bias (WARMSB), weighted average RMSE (WARMSE), and the weighted average standard error of equating (WASE) were calculated to summarize the amount of error over the entire score scale, as shown in Equations 5, 6, and 7 below:

$$WARMSB = \sqrt{\sum_i w_i Bias_i^2}, \tag{5}$$

$$WASE = \sqrt{\sum_i w_i CSE_i^2}, \tag{6}$$

and

$$WARMSE = \sqrt{\sum_i w_i RMSE_i^2}. \tag{7}$$

In Equations 5 through 7, bias, CSE, and RMSE are from Equations 2, 3, and 4. As described previously, *i* is a score point. w_i is the proportion of examinees scoring at each new form score. Weighted statistics were used because the number of examinees scoring at each score point was not the same. Typically, there were a number of score points where no examinees scored.

Comparison across operational and pseudo-test forms. The final research question addressed the extent to which operational and pseudo-test form analyses yielded results that led to the same conclusions. The primary focus of comparison across operational test form and pseudo-test form analyses was at the general findings level.

Results

The results section is arranged into four sections: dimensionality assessment, English Language results, Spanish Language results, and Chemistry results. All operational and pseudo-test form samples for English Language and Spanish Language contained 1,900 examinees. The Chemistry operational and pseudo-test form samples contained 1,500 examinees.

Dimensionality Assessment

Two methods were considered for assessing the dimensionality of the tests analyzed in this study: disattenuated MC and FR correlations and principal components analysis (PCA). For each of the three tests, results are presented for the MC and FR observed and disattenuated correlations as well as for the PCA. Table 5 contains observed and disattenuated correlations, using Cronbach's α as the estimate of reliability. Disattenuated correlations near one would indicate the MC and FR items were measuring the same constructs. For English Language operational test forms, for both test forms and across all sampling conditions, observed correlations were in the range of 0.55 to 0.62. Disattenuated correlations ranged from approximately 0.75 to 0.80. The correlations for the pseudo-test forms tended to be somewhat smaller in magnitude than the correlations for the operational test forms. For the Spanish Language operational test forms, observed correlations were in the range of 0.71 to 0.83. Disattenuated correlations ranged from approximately 0.80 to 0.91. The New Spanish Language form contained 15 fewer MC items; consequently, correlations were lower as compared to the Old form. The correlations for the pseudo-test forms were slightly lower in magnitude to the correlations for the operational test forms. For both Chemistry operational and pseudo-test forms, observed correlations were in the range of 0.85 to 0.88. Disattenuated correlations ranged from approximately 0.94 to 0.95, indicating MC and FR items were measuring essentially the same content and/or processes.

PCA using tetrachoric and polychoric correlations was the second method used to assess the dimensionality of the tests. Scree plots are shown in Figures 1 through 3 for English Language, Spanish Language, and Chemistry, respectively. Across all scree plots, although there were a number of eigenvalues greater than one, it is evident that the first eigenvalue was much larger than the other eigenvalues. The eigenvalues did not level off until after the third or fourth eigenvalues, suggested additional possible weak dimensions. For Spanish Language, the first and second eigenvalues were substantially larger than the other eigenvalues. For English Language, the first principal components accounted for approximately 28% of the total variance, and the ratio of the differences was much larger than three. For Spanish Language, the first principal component accounted for approximately 26% of the total variance, and the ratio of the differences was also larger than three. For Chemistry, the first principal component accounted for approximately 35% of the total variance, and the ratio of the differences was again much larger than three. The combination of evidence suggests that the tests are sufficiently unidimensional for unidimensional IRT analyses because in each case there was a dominant dimension. The same analyses were conducted for the sampling conditions as well as for the pseudo-test forms, and similar results were found but are not shown.

Equating Results: English Language

Descriptive statistics. Descriptive statistics were provided for the original operational English Language 2004 and 2007 test forms in Table 2 and for the pseudo-test forms in Table 4. Effect sizes for the English Language operational test and pseudo-test form samples are shown in Table 6. The effect sizes provided in Table 6 were calculated as new form mean minus old form mean; consequently, a negative effect size indicates that means on the new form were lower than means on the old form. Each row in Table 6 contains effect sizes for the scores listed in the “Score Type” column. The effect sizes are calculated based on the initial single sample of examinees selected at each ES level. Each column contains effect sizes for a given sampling condition. Recall that the target CI ES were 0.00, 0.20, and 0.40 for the operational test forms and 0.00, 0.20, 0.40, and 0.60 for the pseudo-test forms. For the pseudo-test forms, there are two sets of descriptive statistics for the common items: NFR and Full CI. NFR is the common-item set containing only MC items and Full CI is all items in common across the old and new pseudo-test forms. Examinees were sampled using Full CI to determine the target effect sizes. The actual CI ES were within 0.03 of the target CI ES.

Equating relationships. Figure 4 contains the criterion equating relationships for English Language operational and pseudo-test forms. The operational test form plot on the left contains four lines, one for each equating method. The two traditional equating methods (FE and

CE) are illustrated by solid and dashed lines, respectively. The two IRT methods (TS and OS) are illustrated by dotted and dotted-dashed lines, respectively. FE and CE resulted in nearly identical equating relationships, and TS and OS also resulted in nearly identical equating relationships. In addition, the traditional and IRT equating relationships differed by approximately 0.50 score points or less throughout the score scale.

For the pseudo-test forms, the criterion equating relationships were based on a single group equating design for the entire sample of examinees taking the pseudo-test forms. Three criterion equating relationships were calculated as shown on the right of Figure 4: equipercentile, TS, and OS. For both the FE and CE methods, the criterion equating relationship was the equipercentile equating relationship (solid line). For TS and OS, the criterion equating relationships were TS (dotted line) and OS (dotted-dashed line) equating relationships, respectively. It is evident that the single-group equating relationships were very similar.

Conditional bias. Recall that the criterion equating relationship was CI 0.00 for the operational test forms. Figure 5 contains two plots of conditional bias, one for CI 0.20 (left) and one for CI 0.40 (right). There are four lines in each plot, one for each of the four equating methods. The two traditional methods (FE and CE) are illustrated by the solid and dashed lines, respectively. The two IRT methods (TS and OS) are illustrated by the dotted and dotted-dashed lines, respectively. It is evident that as the CI ES increased, conditional bias also increased for FE and CE, though the increase in bias was larger for FE. For both CI 0.20 and CI 0.40, bias for TS and OS was less than 1 score point across the score scale.

Plots of conditional bias for the pseudo-test forms are shown in Figure 6. The top row contains plots of bias for CI 0.00 (left) and CI 0.20 (right). The bottom row contains plots for CI 0.40 (left) and CI 0.60 (right). In each plot, there is one line for each of the four equating methods. FE is represented by the solid line, CE is represented by the dashed line, TS is represented by the dotted line, and OS is represented by the dotted-dashed line. It is evident that for CI 0.00, all four equating methods resulted in similar bias across the score scale. As CI ES increased, bias also increased, though to a greater extent for CE and especially FE. The increase in bias was minimal for TS and OS.

Overall summary statistics. WARMSB, WASE, and WARMSE, which are overall weighted averages across the score scale, are contained in Table 7 for both the operational and pseudo-test forms. The values on the left of the table are for the operational test forms, and the values on the right are for the pseudo-test forms. Consider the values of WARMSB for the operational and pseudo-test forms. For the pseudo-test CI 0.00 condition, bias was similar across all four equating methods. As CI ES increased, values of WARMSB also increased. However, it is evident that for large CI ES, bias was largest for FE, followed by CE. TS and OS resulted in the least bias. The second block of data contains values of WASE. Values of WASE indicate weighted average standard errors of equating across the score scale. For both operational and pseudo-test forms, WASE were smallest for OS across all sampling conditions (approximately 0.34 to 0.37) and largest for CE across all sampling conditions (approximately 0.46 to 0.52). The third block of data contains values for WARMSE, which is an index of the overall average error across the score scale. For both the operational and pseudo-test forms, the same patterns that were found for WARMSB were also found for WARMSE. Additionally, the magnitude of WARMSE appeared to be primarily attributable to bias.

Equating Results: Spanish Language

Descriptive statistics. Descriptive statistics for the operational Spanish Language test forms were provided in Table 2 and for the pseudo-test forms in Table 4. The Spanish Language 2004 test form contained 15 more MC items than Spanish Language 2006. Consequently, results for Spanish Language reflect a situation in which equating (in the strictest sense) may not be advisable. Effect sizes for the Spanish Language operational test forms are shown in Table 8 for both operational and pseudo-test forms. In contrast to the English Language operational forms, CI ES were positive for the operational Spanish Language test forms. The CI ES, shown in the third row, were within 0.03 of the target effect sizes. As described previously for the English Language pseudo-test forms, for the Spanish Language pseudo-test forms, the CI ES were created by sampling examinees using Full CI to obtain the target effect size. Across all sampling conditions, the effect sizes for Full CI were within 0.035 of the target effect sizes.

Equating relationships. Figure 7 contains plots of the criterion equating relationships for Spanish Language original operational and pseudo-test forms. Recall that for the operational test forms, the criterion was the CI 0.00 sampling condition. Consider the plot on the left for the operational criterion equating relationships. FE (solid line) and CE (dashed line) resulted in similar equating relationships, and TS (dotted line) and OS (dotted-dashed line) also resulted in similar equating relationships. However, the traditional and IRT equating relationships were quite different between scores of approximately 50 and 90. Next, consider the plot on the right for the pseudo-test forms. As previously described, the criterion equating relationships for Spanish Language pseudo-test forms were based on a single-group equating. In the plot on the right of Figure 7, it is evident in that the three single-group equating relationships were very similar across the score scale.

Conditional bias. Figure 8 contains plots of conditional bias for the Spanish Language operational test forms. Bias for CI 0.20 equating relationships (left) was near zero across most of the score scale, especially for TS and OS. For CI 0.40, (right), it is evident that as the CI ES increased, bias also increased for all methods. However, bias was much larger for FE and CE as compared to TS and OS. Figure 9 contains four plots of conditional bias for the pseudo-test forms: one each for CI 0.00, CI 0.20, CI 0.40, and CI 0.60. In each plot, there are four lines, one for each equating method. The equating methods are illustrated by solid (FE), dashed (CE), dotted (TS), and dotted-dashed (OS) lines. FE and CE resulted in similar patterns of conditional bias. TS and OS also resulted in similar patterns of conditional bias. As CI ES increased, bias for TS and OS remained close to zero. However, conditional bias for CE, and especially FE, increased substantially as CI ES increased.

Overall summary statistics. Table 9 contains WARMSB, WASE, and WARMSE for Spanish Language operational and pseudo-test forms. As was previously found for the English Language operational test forms, WARMSB tended to increase as CI ES increased. WARMSB was also substantially lower for TS and OS as compared to FE and CE, especially for the operational test forms. For the pseudo-test forms, WARMSB did not exhibit a consistent increasing trend for TS and OS. Additionally, WARMSB for TS and OS was larger for CI 0.20 than for FE and CE. Values of WASE were much larger for the Spanish Language operational test forms as compared to the English Language operational test forms. This likely resulted from the large number of score points and small number of examinees. Additionally, values of WASE appeared to vary across sampling conditions, particularly for TS and OS. Across sampling conditions, values of WASE ranged from approximately 0.74 to 0.92 score points for TS and OS. In contrast to English Language, WASE was smallest for FE for all but one of the operational CI ES. CE resulted in the largest values of WASE for all conditions. For the pseudo-test forms, OS typically resulted in the smallest values of WASE and CE typically resulted in the largest values of WASE. Patterns of WARMSE were similar to those for WARMSB. However, for the operational test forms, because values of WASE were large in magnitude, WASE contributed to the overall error to a greater extent than what was found for English Language.

Equating Results: Chemistry

Descriptive statistics. Descriptive statistics for the Chemistry operational test forms were provided in Table 2 and in Table 4 for the pseudo-test forms. Chemistry 2007 contained two more points than Chemistry 2005. Effect sizes for the Chemistry operational and pseudo-test forms are shown in Table 10. The effect sizes were calculated as new form minus old form scores; consequently, a negative effect size indicates that scores on the new form were lower than scores on the old form. Similar to Spanish Language operational test forms, CI ES for Chemistry were positive. The CI ES effect sizes were within 0.03 of the target effect sizes.

Equating relationships. Figure 10 contains a comparison of the equating relationships for Chemistry operational and pseudo-test forms. Consider first the plot on the left for the operational test forms. Results for CE and FE were similar to each other, and TS and OS were also similar to each other; though differences did exist between the traditional and IRT equating methods. For the pseudo-test forms, the criterion equating relationships (right) were based on a single group equating for the entire sample of examinees taking the pseudo-test forms. The single-group equating relationships were very similar across most of the score scale.

Conditional bias. Plots of conditional bias for the operational test forms are shown in Figure 11. For CI 0.20 (left), bias was around zero across the score scale, except at low scores for TS and OS and high scores for FE and CE. For CI 0.40 (right), conditional bias appeared to increase, except at low scores for TS and OS. Plots of conditional bias for the pseudo-test forms are shown in Figure 12. The figure contains four plots, one for each of the four sampling conditions. In each plot, there are four lines: one for each of the four equating methods. It is evident that for CI 0.00 (top left), all four equating methods resulted in small bias across the score scale. For CI 0.20 (top right) and CI 0.40 (bottom left), TS and OS appeared to result in less bias than FE and CE. Bias was near 0 across the score scale for TS and OS. For CI 0.60 (bottom right), TS and OS still appeared to result in less bias than FE and CE.

Overall summary statistics. Table 11 contains WARMSB, WASE, and WARMSE for Chemistry operational and pseudo-test forms. For the operational test forms, as previously found for the English Language and Spanish Language operational test forms, as CI ES increased, WARMSB also increased for FE and CE. However, for TS and OS, WARMSB actually decreased as the CI ES increased. Across all four equating methods, WARMSB was higher for TS and OS for CI 0.20, but WARMSB was similar across all equating methods for CI 0.40. For the pseudo-test forms, values of WARMSB tended to increase as the CI ES increased, although WARMSB was slightly lower for TS and OS for CI 0.20. By comparing results across the four equating methods in the WARMSB block of data, it is also evident that TS and OS always resulted in the smallest values of WARMSB. CE resulted in the largest values of WARMSB for CI 0.00, and FE resulted in the largest values of WARMSB for the remaining three CI ES. OS always resulted in the smallest WASE, and CE always resulted in the largest WASE, for both operational and pseudo-test forms. Values of WASE were larger for Chemistry as compared to English Language, and similar to or smaller than for Spanish Language. Patterns for WARMSE were similar to WARMSB. However, because values of WASE were somewhat large in magnitude, WASE contributed to the overall error to a greater extent than for English Language.

Discussion

This study examined tests in three different subject areas (English Language, Spanish Language, and Chemistry) that were presumed to represent different degrees of dimensionality. Further, these tests were comprised of different numbers and types of MC and FR items. Both traditional and IRT equating methods were considered. Last, this study incorporated both operational test forms and pseudo-test forms in order to evaluate how findings may differ across the two classes of data. The discussion is divided into three sections: summary of findings, limitations, and implications for future research.

Summary of Findings

Research questions one and two. As the difference in proficiency between old and new form examinee groups increased, equating relationships tended to become more biased relative to the criterion equating relationship. (The criterion equating relationship represented no difference in proficiency between groups of examinees on the old and new test forms.) However, the increase in bias was not consistent across equating methods or tests. When the common-item effect size was large, bias was typically larger for FE than for CE. Bias was typically smallest for TS and OS. Further, bias did not always increase for TS and OS as the difference in proficiency between old and new form groups of examinees increased. It is important to note, however, that a number of factors may have interacted to contribute to these findings. First, the criterion equating relationship was different for each equating method; consequently, comparisons across methods may not be reasonable. Second, only one smoothing value was selected for all replications for the traditional equating methods. This smoothing value may not have been optimal for all replications, introducing additional bias or random error. Results were similar across operational and pseudo-test forms.

For the most part, the findings in this study regarding group differences confirm findings from previous research. Many studies have found that equating tends to be more accurate when there are only small differences in proficiency between groups of examinees taking the old and new test forms (Cao 2008; Kim & Lee, 2006; Kirkpatrick, 2005; Lee et al., 2012; Wang et al., 2008; Wu et al., 2009). Further, this study confirms findings from previous research that CE may be less sensitive than FE to differences in group proficiency (Lee et al., 2012; Wang et al., 2008). von Davier and Wilson (2008) found that TS and CE performed similarly, and that both equating methods were relatively invariant across groups. However, both Cao (2008) and Kirkpatrick (2005) found that differences in group proficiency impacted equating results in the IRT framework.

Research question two. OS most often resulted in the smallest standard errors and CE most often resulted in the largest standard errors. Occasionally, TS and OS resulted in larger standard errors than the traditional equating methods. Additional investigation is needed to determine what caused this result. Although no previous research was found comparing standard errors of equating for traditional and IRT equating methods, Tsai et al. (2001) found that OS generally resulted in smaller standard errors than TS. Further, a number of studies have found that CE tends to result in larger standard errors of equating than FE (Lee et al., 2012; Sinharay & Holland, 2007; Wang et al., 2008).

Research question three. There is some evidence in this study that higher MC and FR correlations or certain subject areas lead to less bias in equating results. Disattenuated MC and FR correlations were lowest for the English Language test forms and highest for the Chemistry test forms. Although disattenuated MC and FR correlations were lowest for English Language, English Language did not consistently result in the highest values of standardized WARMSB. However, given the unique equating situation presented by the Spanish Language operational test forms, it is not surprising that Spanish Language resulted in larger values of WARMSB. The results provide some evidence that either the subject area or the disattenuated MC and FR correlation impacted equating results. Given the tests investigated in this study, it is impossible to disentangle the influence of subject area from MC and FR correlation. However, Lee et al. (2012) found that equating was more accurate for a common-item set containing only MC items when the correlation between MC and FR scores was higher.

Research question four. The two classes of data resulted in similar conclusions; however, across the three tests, there were some differences in the findings based on operational test forms as compared to pseudo-test forms. For both the English Language operational and pseudo-test form analyses, the conclusions were generally the same. WARMSB was largest for FE and smallest for TS and OS. Additionally, standard errors of equating were similar, although they were smaller in magnitude for the pseudo-test forms. However, for the operational test forms, standard errors of equating were always smallest for OS. For the pseudo-test forms, standard errors of equating were smallest for OS for only half of the sampling conditions.

Results for Spanish Language operational and pseudo-test forms were less similar than those for English Language. The Spanish Language operational test forms represented a unique situation where equating, in the strictest sense, could not be done because the test forms were of unequal lenghth. Values of WASE for the operational test forms were only smallest for OS for one of the three sampling conditions. For the pseudo-test forms, values of WASE were smallest for OS for all of the sampling conditions.

In general, for Chemistry, similar results were seen across operational and pseudo-test forms. For FE and CE, similar patterns of WARMSB were seen across the operational test and pseudo-test forms. WARMSB steadily increased as the common-item effect size increased for the sampling conditions. However, patterns of WARMSB were somewhat different for TS and OS across operational test and pseudo-test forms. For all sampling conditions, WASE was smallest for OS and largest for CE.

Most research has not compared operational and pseudo-test forms. Sinharay and Holland (2006, 2007) used operational test, pseudo-test, and simulated test forms for MC-only tests. Similar to this study, they found that the different classes of data led to similar conclusions.

Limitations

Resampling study. Although sampling was conducted to create various levels of effect sizes, the sampling process also resulted in differences in standard deviations, skewness, and kurtosis. Furthermore, because of the constraint of simultaneously creating different levels of common-item effect sizes and differences in the effect sizes of MC and FR items, it was not possible to hold old form means constant across sampling conditions. Consequently, the situations to which the results can be generalized were also limited.

Confounding of subject area and correlation. A strength of this study was that it incorporated three different subject area tests representing different correlations between MC and FR items. However, one of the limitations of the operational test forms selected for this study was that the MC and FR correlation levels and subject areas were completely confounded. Consequently, it was impossible to disentangle the influence of the subject area test characteristics from the influence of the MC and FR correlations.

Implications for Future Research

Simulation study. A primary limitation of this study was use of the resampling methodology. The complex interactions among effect sizes illustrate the need to hold certain scores and effect sizes constant while manipulating others. A simulation study could also incorporate various levels of MC and FR correlations for a given subject area. Differences in examinee proficiency on MC and FR items could also be simulated for various levels of MC and FR correlations. Also, different levels of MC and FR correlations in combination with levels of composite score and common-item score correlations may be informative.

Resampling considerations. Although a simulation study may be the best approach to further investigate some of the findings in this study, future research could also be conducted with changes to the resampling study. In this study, common-item effect sizes were created by sampling examinees based on all items in common between test forms. However, different methods could also be considered, such as sampling based on MC common items, FR common items, or both MC and FR common items.

Subject area and correlation. One of the limitations of the tests selected for this study was that the MC and FR correlation levels and subject areas for the tests were completely confounded. Through simulation study or selection of additional tests, future research could incorporate multiple levels of MC and FR correlations for a single subject area. However, tests from similar subject areas also often had similar MC and FR correlations. It may not be possible to select tests from the same subject area with different levels of MC and FR correlations.

Equating research data classes. The results from this study suggest that when pseudo-test forms are constructed to be similar to operational data, the results from both equating research data classes yield similar conclusions. Therefore, it is plausible that when pseudo-tests are constructed to create situations that cannot be researched using operational data, the results will be comparable to the operational situation. Future research should also consider the comparison of simulation studies as compared to pseudo-test forms and operational data.

Overall, the results of this study suggest that the test, examinee, and common-item characteristics investigated in this study do impact equating results. Large differences in the proficiency between old and new form examinee groups may result in larger bias among equating relationships. However, the impact on bias of group differences may be influenced by the correlation between MC and FR items or equating method. Specifically, inclusion of FR items in the common-item set may result in smaller bias in certain situations. Further, TS and OS *might* be preferred when group differences in proficiency are large, although this finding may be dependent on the specific factors of investigation and choice of criterion equating relationships in this study. Last, when the correlation between MC and FR items is high, bias may be relatively small, even for large differences in examinee proficiency. However, future research is needed to determine the specific conditions for which these findings can be expected to hold.

References

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes* (CASMA Monograph Number 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Westport, CT: American Council on Education/Praeger.

Divgi, D. R. (1980). *Dimensionality of binary items: Use of a mixed model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

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

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

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195-240.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10,* 35-43.

Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50,* 61-71.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 129-164.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*, 17-43.

Jiao, H. (2004). *Evaluating the dimensionality of the Michigan English Language Assessment Battery. Spaan fellow working papers in second or foreign language assessment: Volume 2.* University of Michigan, Ann Arbor, MI. Retrieved from http://www.lsa.umich.edu/UMICH/eli/Home/Research/Spaan%20Fellowship/pdfs/spaan_working_papers_v2_jiao.pdf.

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

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

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

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

Kolen, M. J. (2003). *POLYEQUATE* [Computer program]. Iowa City: University of Iowa.

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

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen, & W. Lee (Ed.), *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.

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

Muraki, E., & Bock, D. (2002). *PARSCALE 4.1* [Computer program]. Chicago: Scientific Software International, Inc.

Orlando, M. (2004). *Critical issues to address when applying item response theory (IRT) models.* Paper presented at the Conference on Improving Health Outcomes Assessment Based on Modern Measurement Theory and Computerized Adaptive Testing, Bethesda, MD. Retrieved from http://outcomes.cancer.gov/conference/irt/orlando.pdf.

R Development Core Team (2009*). R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York, NY: John Wiley & Sons.

Ricker, K. L., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design.* Technical Report (RR-07-44). Princeton, N.J.: Educational Testing Service.

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

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505-528.

Sinharay, S., & Holland, P. (2006). *The correlation between the scores of a test and anchor test.* ETS Technical Report (RR-06-04). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*, 249-275.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett and W. C. Ward (Eds.). *Construction versus choice in cognitive measurement* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Table 1

Description of Selected AP Tests

Test	Year	MC Items	FR Items	FR Points	Total Points
English Language	2004	53	3	9	80
	2007	52	3	9	79
Spanish Language	2004	90	20, 5, 2	1, 4, 9	148
	2006	75	20, 5, 2	1, 4, 9	133
Chemistry	2005	75	1, 3, 1, 1	8, 9, 10, 14	134
	2007	75	3, 2, 1	9, 10, 14	136

Table 2

Descriptive Statistics for English Language, Spanish Language, and Chemistry after Imputation

Item Format	Descriptive Statistics	English Language		Spanish Language		Chemistry	
		2004	2007	2004	2006	2005	2007
	N[a]	20,000	20,000	20,000	20,000	20,000	20,000
	N[b]	15,820	16,882	19,010	18,022	13,027	12,328
MC	Mean	37.155	35.201	60.561	50.656	46.962	44.540
	SD	9.350	8.774	15.148	11.908	15.861	15.403
	Skewness	-0.620	-0.540	-0.479	-0.425	-0.430	-0.296
	Kurtosis	2.707	2.773	2.625	2.627	2.132	2.056
	α	0.900	0.886	0.933	0.912	0.949	0.944
FR	Mean	14.692	14.334	37.308	38.598	31.401	29.179
	SD	3.899	3.942	10.547	11.208	13.845	15.601
	Skewness	-0.211	-0.236	-0.669	-0.800	-0.391	-0.198
	Kurtosis	3.122	3.206	2.775	2.967	2.266	2.001
	α	0.636	0.671	0.860	0.872	0.883	0.910
CI	Mean	13.850	13.824	16.927	16.652	15.824	15.801
	SD	3.147	3.089	4.919	4.980	5.579	5.539
	Skewness	-1.075	-1.035	-0.332	-0.255	-0.451	-0.448
	Kurtosis	3.987	3.958	2.445	2.348	2.226	2.199
	CO Corr.	0.834	0.840	0.838	0.847	0.924	0.923
CO	Mean	51.847	49.536	97.869	89.253	78.363	73.719
	SD	12.009	11.511	24.470	21.400	28.696	30.089
	Skewness	-0.568	-0.512	-0.589	-0.679	-0.429	-0.266
	Kurtosis	2.852	2.957	2.751	2.998	2.217	2.033
	α	0.884	0.875	0.947	0.934	0.931	0.929

[a] Before imputation
[b] After imputation

Table 3

Composition of Pseudo-Test Forms

Form	Test	Item Type	Number of MC (Points)	Number of FR (Points)	Difficulty		Discrimination	
					Mean	SD	Mean	SD
Old	English Language	Total Test	36 (36)	2 (18)	0.713	0.131	0.375	0.097
		NFR	17 (17)	0 (0)	0.756	0.124	0.358	0.124
		Full CI	19 (19)	1 (9)	0.756	0.103	0.375	0.102
	Spanish Language	Total Test	68 (68)	20 (40)	0.648	0.177	0.376	0.124
		NFR	33 (33)	0 (0)	0.649	0.151	0.358	0.097
		Full CI	46 (46)	13 (22)	0.631	0.169	0.370	0.118
	Chemistry	Total Test	54 (54)	4 (42)	0.588	0.155	0.464	0.139
		NFR	28 (28)	0 (0)	0.641	0.155	0.460	0.113
		Full CI	33 (33)	2 (23)	0.617	0.151	0.475	0.121
New	English Language	Total Test	36 (36)	2 (18)	0.706	0.121	0.402	0.101
		NFR	17 (17)	0 (0)	0.756	0.124	0.364	0.119
		Full CI	19 (19)	1 (9)	0.756	0.103	0.380	0.099
	Spanish Language	Total Test	68 (68)	20 (40)	0.641	0.167	0.382	0.136
		NFR	33 (33)	0 (0)	0.649	0.151	0.350	0.107
		Full CI	46 (46)	13 (22)	0.631	0.169	0.360	0.135
	Chemistry	Total Test	54 (54)	4 (42)	0.601	0.152	0.465	0.128
		NFR	28 (28)	0 (0)	0.641	0.155	0.459	0.115
		Full CI	33 (33)	2 (23)	0.617	0.151	0.475	0.122

Table 4

Descriptive Statistics for English Language, Spanish Language, and Chemistry Pseudo-Test Forms

Item Format	Descriptive Statistics	English Language		Spanish Language		Chemistry	
		Old	New	Old	New	Old	New
	N	15,820	15,820	19,010	19,010	12,328	12,328
MC	Mean	26.023	25.735	45.029	44.486	32.159	32.898
	SD	6.042	6.602	11.286	11.795	11.341	11.361
	Skewness	-0.713	-0.774	-0.401	-0.398	-0.332	-0.353
	Kurtosis	3.003	3.036	2.624	2.541	2.094	2.086
	α	0.841	0.867	0.909	0.914	0.926	0.927
FR	Mean	9.564	9.760	25.438	27.110	20.765	20.894
	SD	2.842	2.805	7.030	7.930	10.951	11.269
	Skewness	-0.208	-0.122	-0.614	-0.732	-0.255	-0.233
	Kurtosis	3.075	2.977	2.897	2.662	2.027	1.986
	α	0.575	0.527	0.821	0.818	0.864	0.868
NFR	Mean	12.853	12.853	21.423	21.423	17.941	17.941
	SD	3.042	3.042	5.855	5.855	6.300	6.300
	Skewness	-0.991	-0.991	-0.321	-0.321	-0.521	-0.521
	Kurtosis	3.726	3.726	2.485	2.485	2.229	2.229
	CO Corr.	0.846	0.859	0.913	0.891	0.936	0.936
Full CI	Mean	19.236	19.236	44.194	44.194	32.997	32.997
	SD	4.216	4.216	11.359	11.359	13.242	13.242
	Skewness	-0.841	-0.841	-0.455	-0.455	-0.431	-0.431
	Kurtosis	3.656	3.656	2.700	2.700	2.117	2.117
	CO Corr.	0.918	0.922	0.983	0.969	0.984	0.983
CO	Mean	35.587	35.496	70.467	71.596	52.923	53.792
	SD	7.848	8.378	17.245	18.536	21.498	21.799
	Skewness	-0.628	-0.654	-0.517	-0.575	-0.316	-0.318
	Kurtosis	3.069	3.031	2.809	2.677	2.064	2.047
	α	0.820	0.842	0.929	0.930	0.899	0.898

Table 5

Observed and Disattenuated MC and FR Correlations

Test	Form	Corr.	Operational				Pseudo-Test				
			Total	CI 0.00	CI 0.20	CI 0.40	Single Group	CI 0.00	CI 0.20	CI 0.40	CI 0.60
English Lang.	Old	MC/FR	0.571	0.617	0.605	0.584	0.495	0.487	0.495	0.484	0.479
		Dis. MC/FR	0.755	0.769	0.763	0.737	0.712	0.687	0.708	0.690	0.694
	New	MC/FR	0.578	0.570	0.559	0.583	0.506	0.505	0.542	0.519	0.521
		Dis. MC/FR	0.750	0.754	0.749	0.790	0.749	0.736	0.765	0.752	0.751
Spanish Lang.	Old	MC/FR	0.808	0.790	0.770	0.827	0.760	0.781	0.779	0.786	0.730
		Dis. MC/FR	0.902	0.889	0.873	0.914	0.880	0.893	0.893	0.896	0.864
	New	MC/FR	0.714	0.746	0.768	0.752	0.757	0.747	0.762	0.791	0.752
		Dis. MC/FR	0.800	0.833	0.844	0.836	0.875	0.860	0.874	0.899	0.869
Chemistry	Old	MC/FR	0.866	0.859	0.858	0.862	0.860	0.864	0.855	0.860	0.862
		Dis. MC/FR	0.946	0.939	0.941	0.947	0.961	0.960	0.956	0.952	0.958
	New	MC/FR	0.884	0.876	0.870	0.867	0.856	0.866	0.848	0.860	0.860
		Dis. MC/FR	0.953	0.945	0.946	0.943	0.954	0.959	0.950	0.959	0.956

Table 6

Effect Sizes for English Language Test Forms

	Operational Test Forms			Pseudo-Test Forms				
Score Type	CI 0.00	CI 0.20	CI 0.40	Single Group	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	1,900	1,900	1,900	15,820	1,900	1,900	1,900	1,900
NFR CI	-0.014	-0.204	-0.413	0.000	-0.014	-0.183	-0.358	-0.495
Full CI	--	--	--	0.000	-0.021	-0.229	-0.396	-0.582

Table 7

Summary Statistics for English Language Test Forms

Statistic	Equating Method	Operational Test Forms			Pseudo-Test Forms			
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	--	1.489	2.324	0.192	0.856	1.111	1.819
	CE	--	1.189	1.612	0.198	0.632	0.713	1.327
	TS	--	0.679	0.658	0.197	0.404	0.355	0.632
	OS	--	0.672	0.653	0.139	0.413	0.385	0.682
WASE	FE	0.408	0.427	0.445	0.279	0.275	0.287	0.311
	CE	0.467	0.522	0.519	0.311	0.319	0.323	0.343
	TS	0.358	0.380	0.367	0.252	0.243	0.252	0.270
	OS	0.341	0.361	0.348	0.234	0.219	0.234	0.246
WARMSE	FE	--	1.549	2.366	0.338	0.899	1.147	1.845
	CE	--	1.298	1.694	0.368	0.708	0.783	1.371
	TS	--	0.778	0.753	0.320	0.472	0.435	0.688
	OS	--	0.763	0.740	0.272	0.467	0.451	0.725

Table 8

Effect Sizes for Spanish Language Test Forms

	Operational Test Forms			Pseudo-Test Forms				
ES	CI 0.00	CI 0.20	CI 0.40	Single Group	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	1,900	1,900	1,900	19,010	1,900	1,900	1,900	1,900
CO	-0.468	-0.223	0.223	0.063	0.096	-0.114	-0.366	-0.616
NFR CI	-0.008	0.216	0.423	0.000	0.023	-0.139	-0.346	-0.590
Full CI	--	--	--	0.000	0.009	-0.181	-0.378	-0.633
MC	-0.783	-0.561	-0.101	-0.047	-0.008	-0.214	-0.444	-0.689
FR	-0.007	0.227	0.602	0.223	0.242	0.059	-0.193	-0.427

Table 9

Summary Statistics for Spanish Language Test Forms

Statistic	Equating Method	Operational Test Forms			Pseudo-Test Forms			
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	--	2.365	6.639	1.033	1.146	1.738	2.542
	CE	--	1.601	5.069	0.904	0.929	1.219	1.576
	TS	--	0.435	2.113	0.493	0.519	0.623	0.416
	OS	--	0.397	2.064	0.485	0.463	0.651	0.431
WASE	FE	0.870	0.847	0.836	0.521	0.555	0.554	0.623
	CE	1.022	1.011	0.969	0.615	0.654	0.646	0.708
	TS	0.912	0.921	0.750	0.467	0.531	0.521	0.437
	OS	0.892	0.907	0.737	0.450	0.514	0.505	0.423
WARMSE	FE	--	2.512	6.691	1.157	1.274	1.824	2.617
	CE	--	1.893	5.161	1.093	1.136	1.379	1.728
	TS	--	1.018	2.242	0.679	0.743	0.812	0.604
	OS	--	0.990	2.191	0.662	0.691	0.824	0.604

Table 10

Effect Sizes for Chemistry Test Forms

	Operational Test Forms			Pseudo-Test Forms				
ES	CI 0.00	CI 0.20	CI 0.40	Single Group	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	1,900	1,900	1,900	12,328	1,500	1,500	1,500	1,500
CO	-0.120	0.043	0.272	0.040	0.039	0.285	0.464	0.648
NFR CI	0.012	0.183	0.422	0.000	-0.014	0.224	0.398	0.585
Full CI	--	--	--	0.000	-0.005	0.228	0.415	0.599
MC	-0.118	0.040	0.264	0.065	0.059	0.295	0.468	0.651
FR	-0.113	0.044	0.261	0.012	0.015	0.252	0.427	0.599

Table 11

Summary Statistics for Chemistry 2005-2007 Operational Test Forms

Statistic	Equating Method	Operational Test Forms			Pseudo-Test Forms			
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	--	0.790	1.313	0.486	0.928	1.304	1.598
	CE	--	1.128	1.214	0.646	0.697	0.819	1.032
	TS	--	1.635	1.269	0.365	0.219	0.436	0.627
	OS	--	1.569	1.254	0.363	0.204	0.419	0.593
WASE	FE	0.904	0.866	0.843	0.610	0.596	0.598	0.636
	CE	1.029	0.983	0.970	0.683	0.677	0.698	0.716
	TS	0.726	0.821	0.616	0.422	0.441	0.477	0.525
	OS	0.715	0.814	0.607	0.415	0.435	0.472	0.520
WARMSE	FE	--	1.172	1.560	0.780	1.103	1.435	1.720
	CE	--	1.496	1.554	0.940	0.972	1.076	1.256
	TS	--	1.830	1.410	0.558	0.493	0.646	0.818
	OS	--	1.767	1.394	0.551	0.481	0.631	0.789

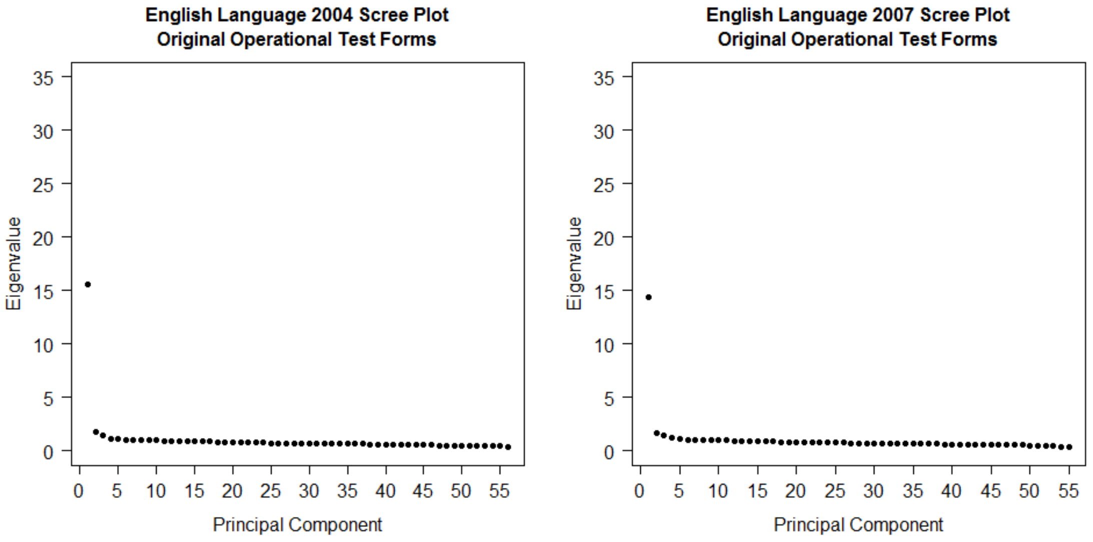


Figure 1. English Language operational test form scree plots.

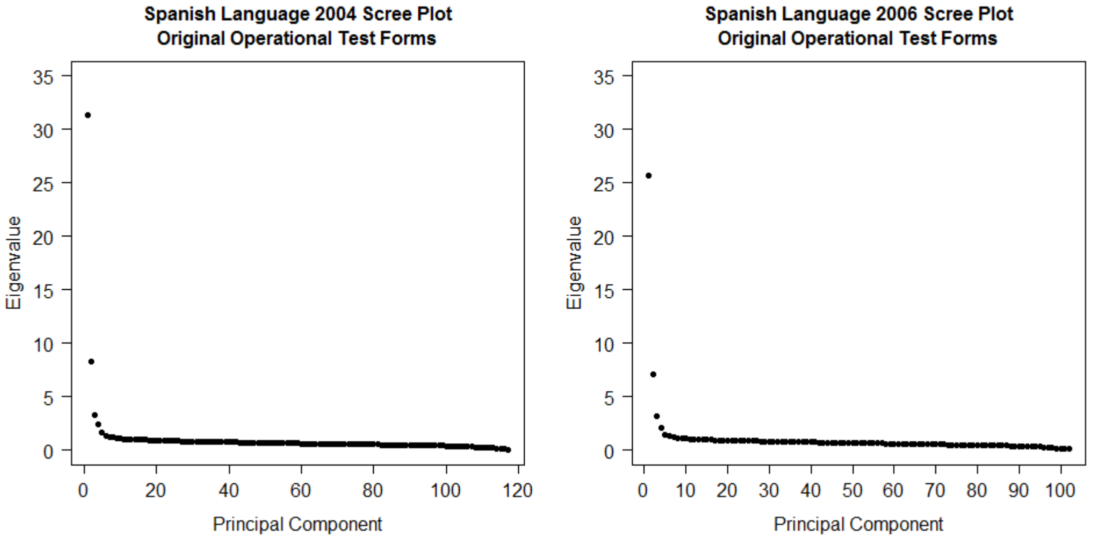


Figure 2. Spanish Language operational test form scree plots.

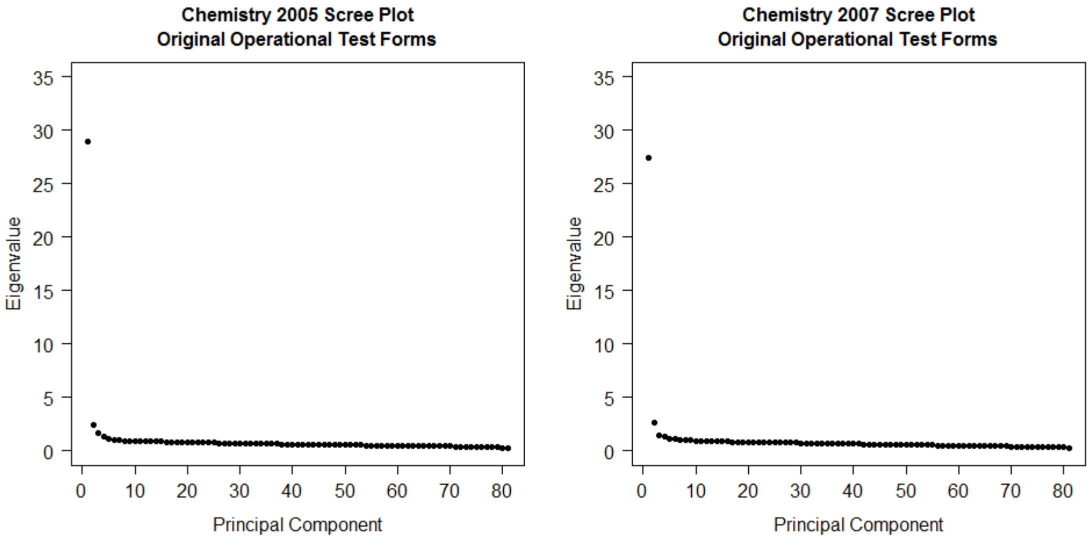


Figure 3. Chemistry operational test form scree plots.

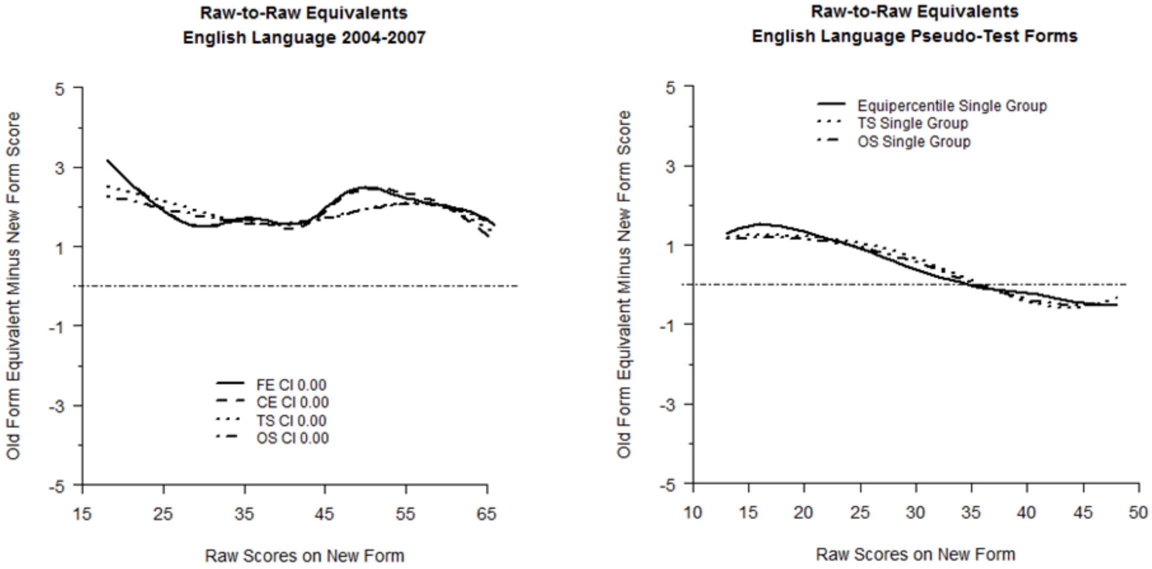


Figure 4. Criterion equating relationships for English Language operational and pseudo-test forms.

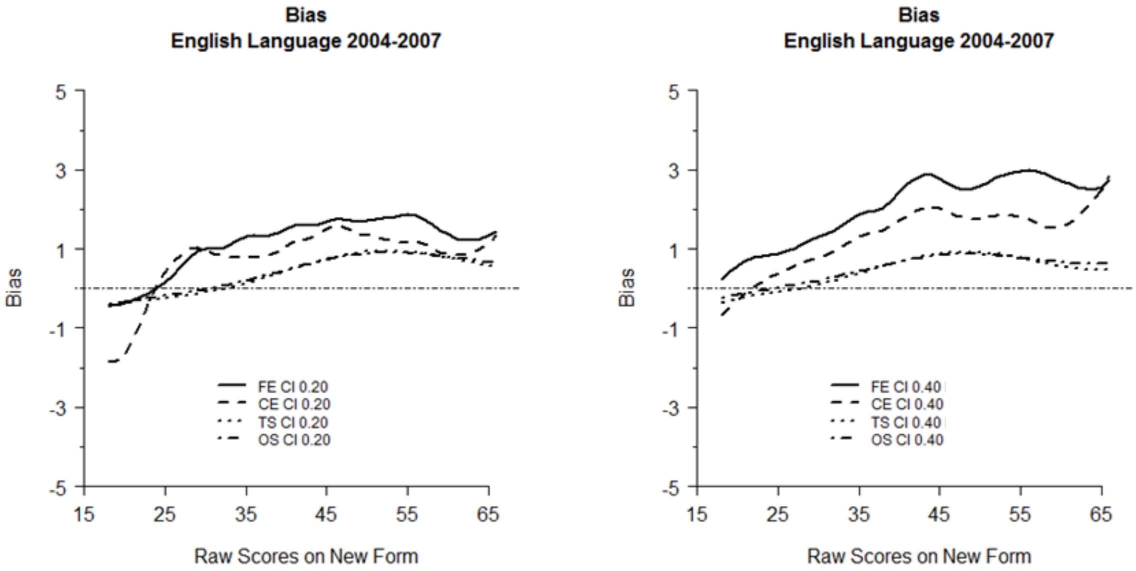


Figure 5. Conditional bias for English Language operational test forms.

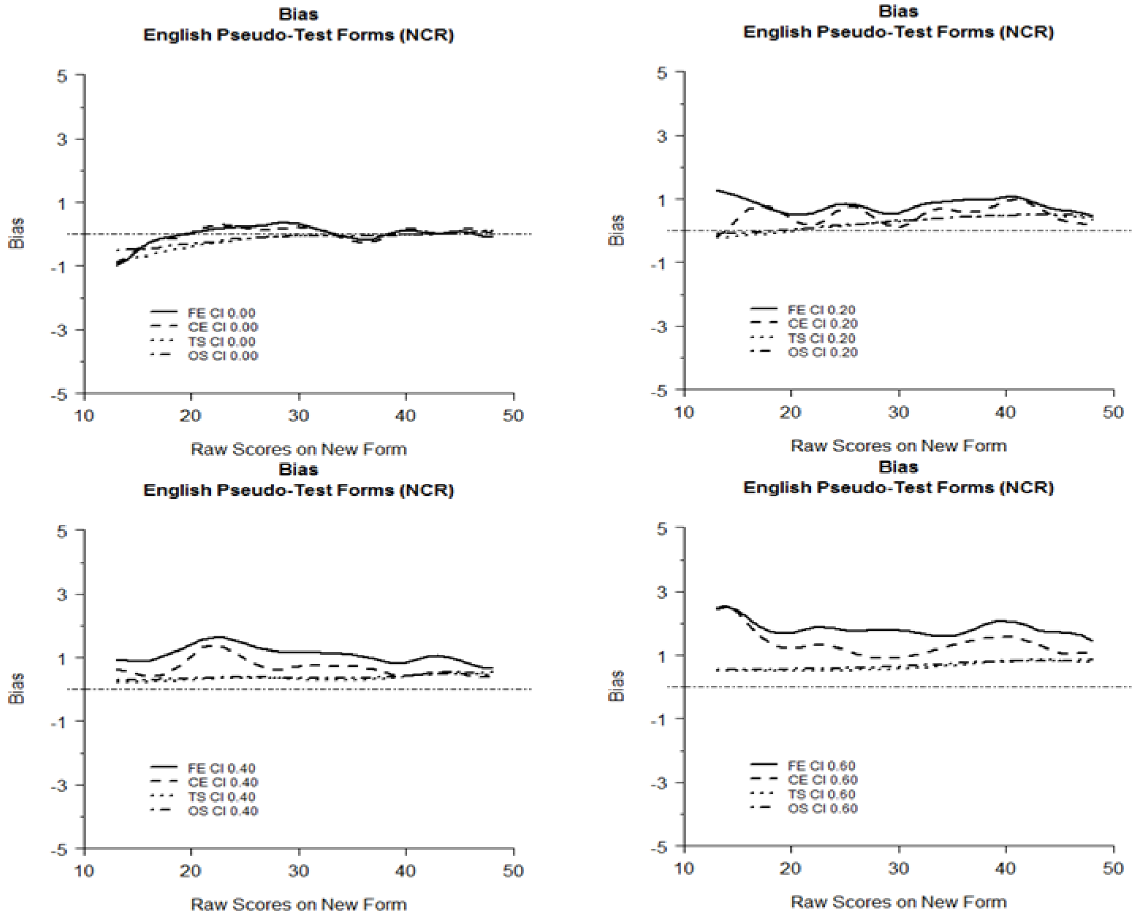


Figure 6. Conditional bias for English Language pseudo-test forms.

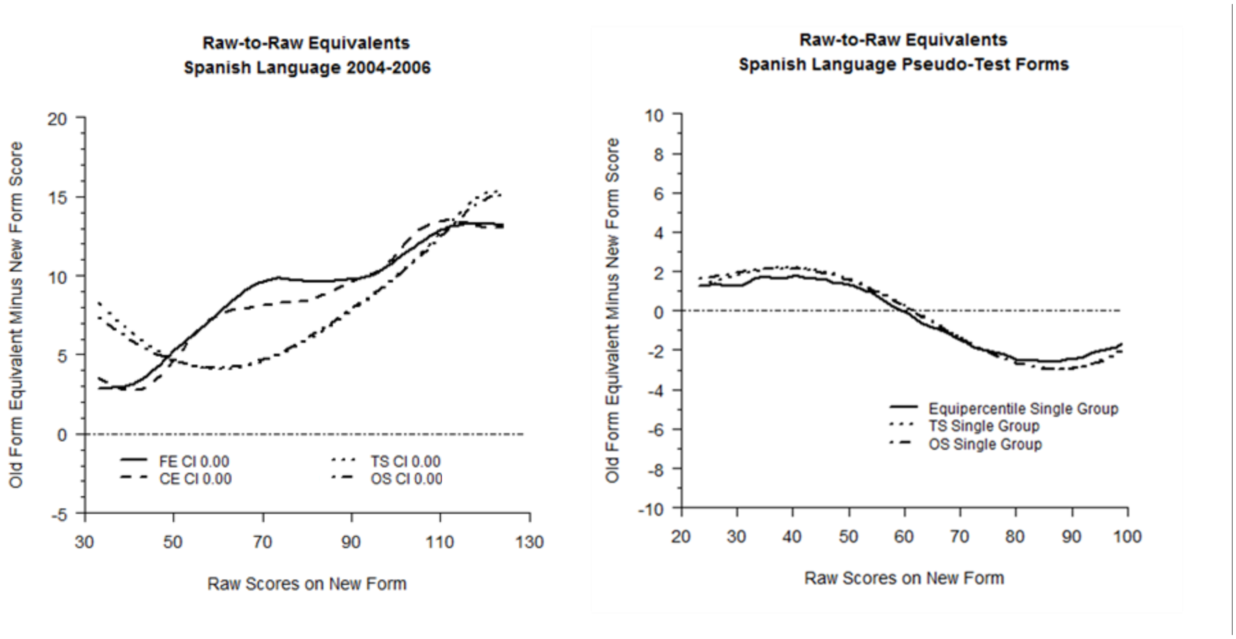


Figure 7. Criterion equating relationships for Spanish Language operational and pseudo-test forms.

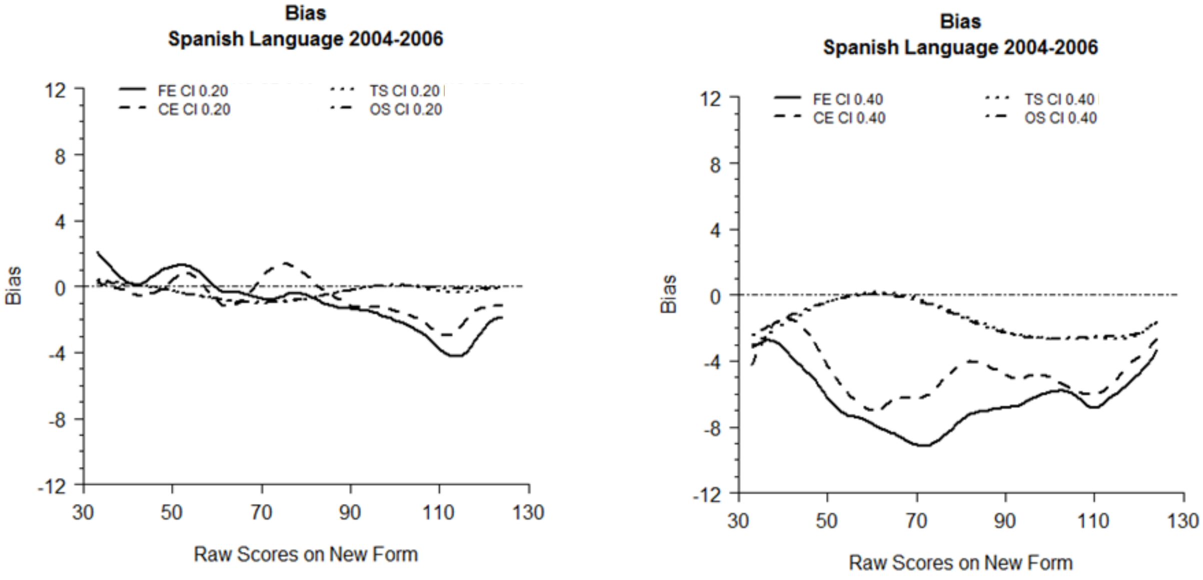


Figure 8. Conditional bias for Spanish Language operational test forms.

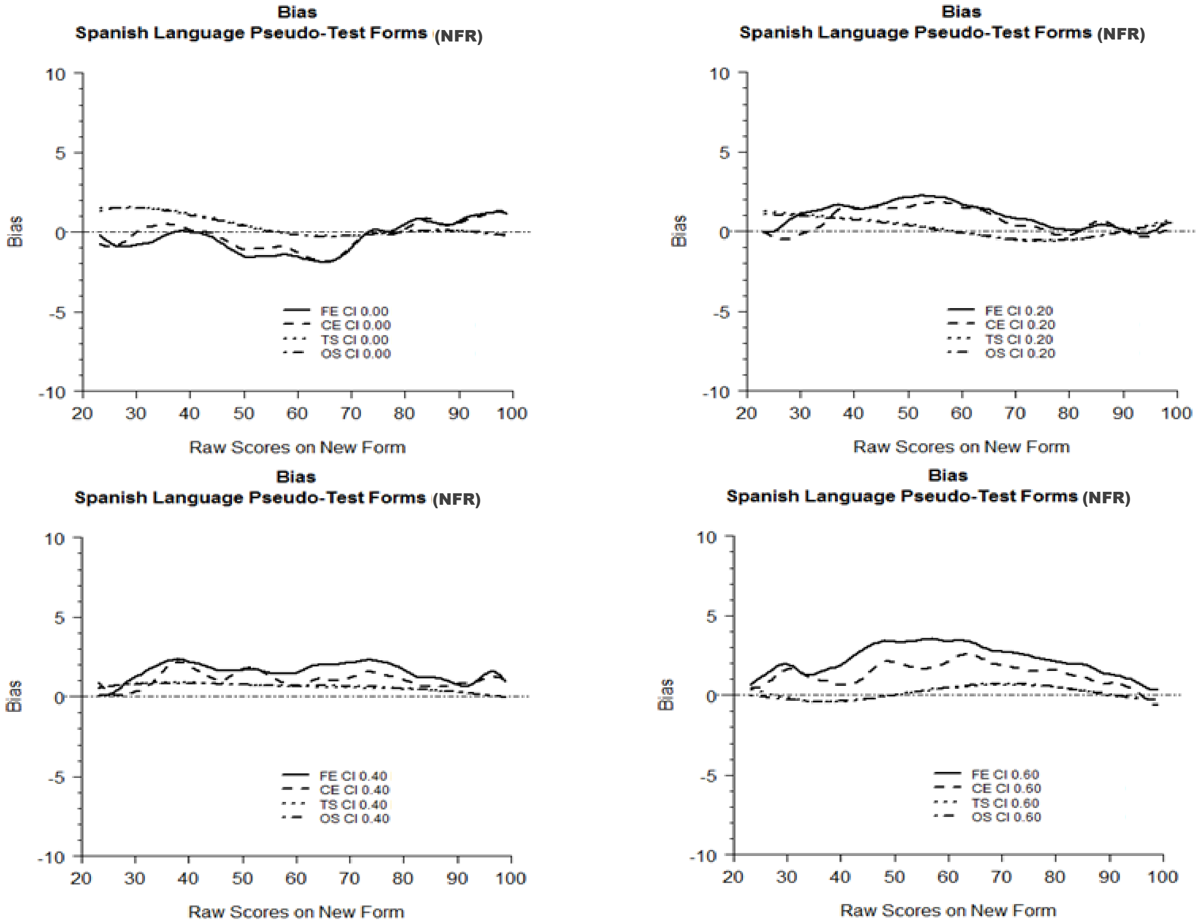


Figure 9. Conditional bias for Spanish Language pseudo-test forms.

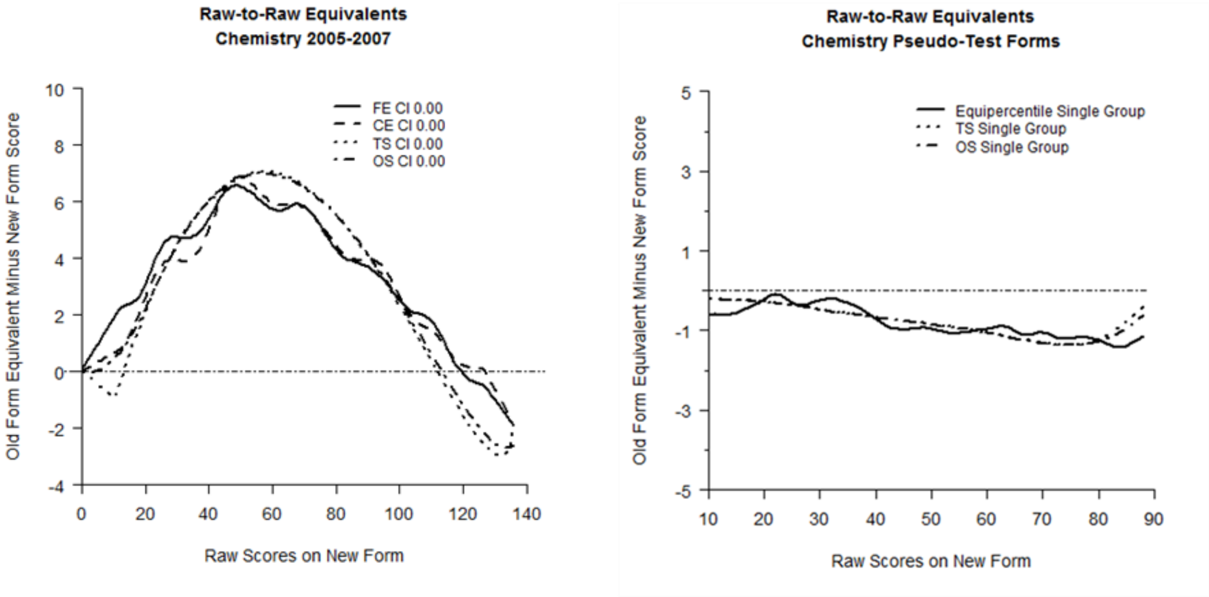


Figure 10. Criterion equating relationships for Chemistry operational and pseudo-test forms.

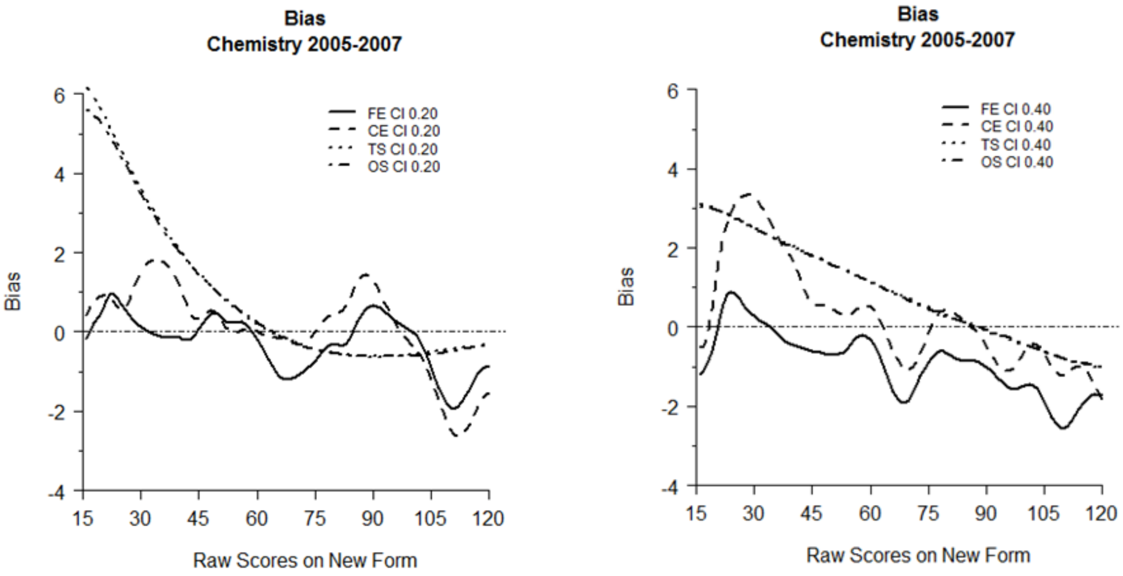


Figure 11. Conditional bias for Chemistry operational test forms.

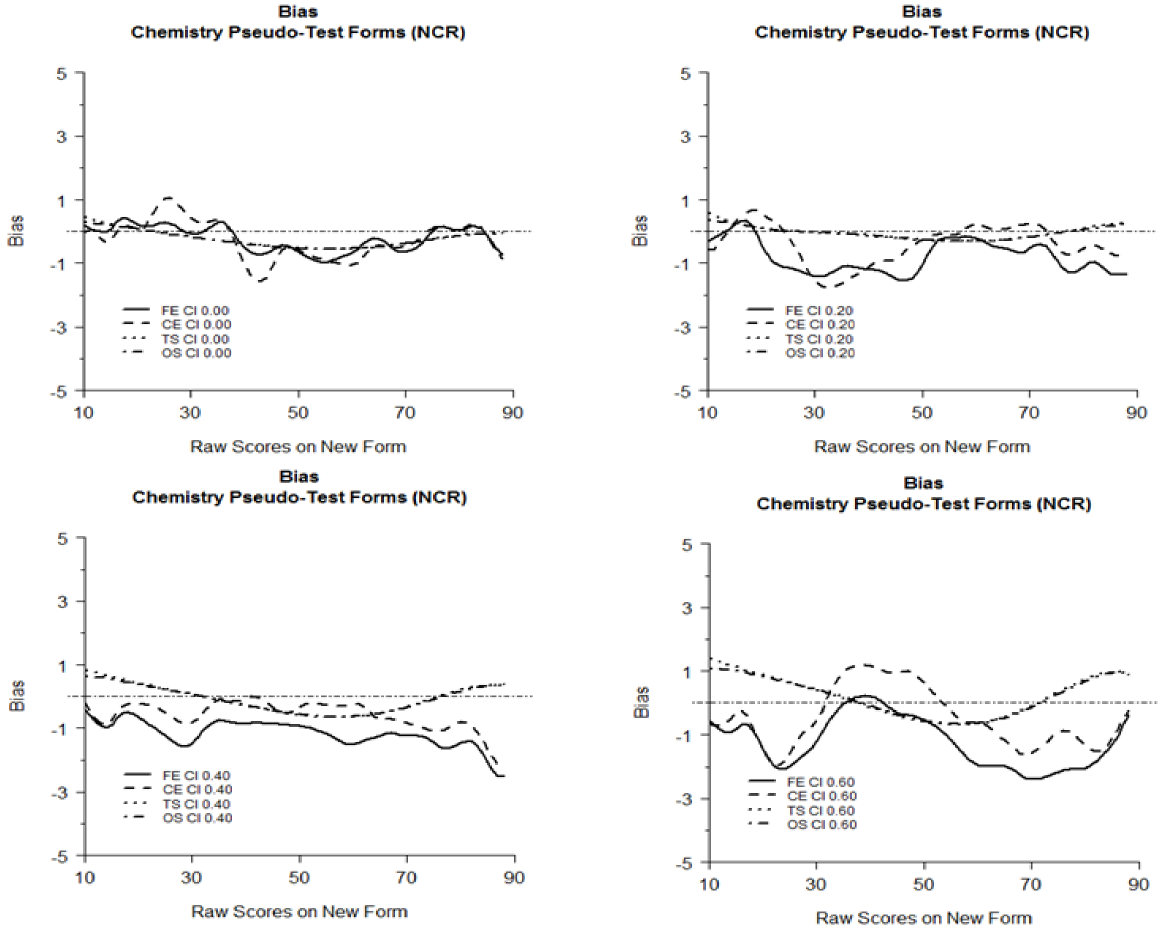


Figure 12. Conditional bias for Chemistry pseudo-test forms.

Chapter 4: Using Matched Samples Equating Methods to Improve Equating Accuracy

Sonya Powers and Michael J. Kolen

The University of Iowa, Iowa City, IA

Abstract

Research indicates that equating results become less accurate as group differences increase. In this study, matched samples equating methods were considered as a way to decrease group differences in order to improve equating results. Group mean differences on a mixed-format exam were manipulated to range from 0 to 0.75 standard deviation units in performance on common items. Equating accuracy was evaluated across four curvilinear equating methods and four matching methods. As group differences increased, equating results became increasingly biased and dissimilar across equating methods. Matching on the selection variable, with or without other variables, provided more accurate equating results compared to equating using unmatched groups. Matching on variables only moderately related to the selection variable did not greatly improve equating accuracy. The frequency estimation equipercentile equating method appeared to be the most sensitive to group differences, but benefitted greatly from matching on the selection variable. Large group differences appeared to violate the assumptions of the frequency estimation and chained equipercentile equating methods. Matching on the selection variable appeared to greatly improve the degree to which the assumptions held for these two methods. A relationship between equation accuracy and the degree to which IRT assumptions held was not identified.

Using Matched Samples Equating Methods to Improve Equating Accuracy

Equating methods are used to adjust scores on exam forms for differences in difficulty. The common-item nonequivalent group (CINEG) equating design (sometimes referred to the nonequivalent anchor test or NEAT design) has been used in situations where groups of examinees with different average ability take forms of varying difficulty. The CINEG design provides a way to adjust for group differences and form differences by including a subset of items from a previous form into a new form. An assumption of the equating methods used with this design is that performance differences on the common item set indicate group differences that generalize to the rest of the items on the form. The estimate of group differences based on the common items is used so that score adjustments are made based on differences in form difficulty, and not based on differences in group performance.

Several equipercentile and item response theory (IRT) equating methods have been developed for use with the CINEG design. According to Kolen (1990), when group differences are fairly small and exam forms and common items are constructed to be nearly parallel in terms of content and statistical properties, all equating methods tend to give reasonable and similar results. Many empirical studies provide evidence in support of Kolen's conclusion (e.g., Cook, Dunbar, & Eignor, 1981; Cook, Eignor, & Taft, 1998). However, research has also indicated that various equating methods provide different score adjustments when group differences are large (e.g., Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Stocking & Eignor, 1986). Powers and Kolen (2011) found that large group differences can result in violations of the assumptions of CINEG equating methods which may explain the inaccurate equating results. As noted by Kolen (1990), "if the common items do not behave in the same way in the old and new groups, then no equating method can be expected to function adequately."

Because equating with large group differences may provide inaccurate and inconsistent equating results, a possible solution is to equate with more similar groups. One technique for creating more similar groups across forms is to match the group taking each form based on variables related to ability. This technique is called matched samples equating. If matched samples equating reduces the extent to which equating assumptions are violated, then the accuracy of equating results may be improved.

Matching has historically been used in experimental design to increase the power of statistical testing by decreasing random error. It has also been used in quasi-experimental situations when there is no a priori way to ensure that treatment and control groups are randomly equivalent. More recently, the use of matching nonequivalent groups based on background variables or test scores has been used in psychometric research when nonequivalent groups take different forms (or formats) of an exam.

Researchers at the Educational Testing Service conducted a series of studies to evaluate the impact of matched samples equating on equating accuracy and consistency (Cook, Eignor, & Schmitt, 1988, 1990; Lawrence & Dorans, 1990; Livingston, Dorans & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990). They used data from a variety of large-scale assessments and evaluated several equating methods including Levine true and observed score equating, Tucker, frequency estimation, chained equipercentile, and IRT true score equating. They also used several different criteria for evaluating equating results including the consistency of equating results across methods, the similarity of matched samples equating relationships to an equating relationship developed with similar examinee groups, and equating a test to itself. Matching variables included common item scores and self-reported responses to questions about examinee ability in the content area. A consistent finding across all studies was that matched samples equating results tended to result in more similar equating results across equating methods. In some studies the matched equating results were closer to the criterion than unmatched results (Cook et al., 1988, , 1990; Schmitt et al., 1990). In other studies, the matched equating results appeared more biased than unmatched results (Eignor, Stocking, & Cook, 1990; Livingston et al., 1990; see also a more recent study by Paek, Liu, & Oh, 2006). Eignor et al. (1990) and Livingston et al. (1990) suggested that matching on common item scores causes biased equating results and should not be used. Matching on self-report variables did not improve equating accuracy in any of the studies where they were included. Also, there was no consistency across studies in the equating methods that were found to be most sensitive to group differences. Skaggs (1990) noted that a limitation of the ETS matched samples equating studies was that they did not provide the standard error of equating (SE), so it was difficult to tell which differences between equating methods and sampling procedures were relevant, and which were within sampling error.

Wright and Dorans (1993) manipulated group differences using examinee scores on a different but correlated measure. In this study, the ‘selection variable,’ or the variable that groups differed on was known. They then used scores on the correlated measure to match the old form and new form groups. Using this matching procedure, they were able to obtain more accurate equating results than when equating with the unmatched groups. This finding indicates that matching on the selection variable(s), if possible, holds promise as a method of improving the accuracy of equating results.

Matched sampling has also been studied as a way to adjust for mode effects when forms are administered in paper and computerized formats. In these studies, the items taken by both groups are the same but the groups self-select into administration mode and differ in ability and a variety of other characteristics. Although the items are the same, there is concern that administration mode may affect item difficulty because of issues like speededness, computer skills, and differences in the way items are displayed. In order to draw conclusions about mode effects, it is necessary to eliminate the confounding influence of group differences. Matching samples on background variables has been investigated as a possible solution to eliminate the confounding of group and mode differences.

In one such study, Yu, Livingston, Larkin, and Bonett (2004) used matched sampling with the Pre-Professional Skills Test (PPST) to try to compare a paper-based administration group to a computerized administration group that differed by approximately half a standard deviation. The authors used logistic regression to assign propensity scores (Rosenbaum & Rubin, 1983) to examinees. Matching variables included gender, ethnicity, educational background, job related information, and teaching experience. After matching, the mean performance difference between the two groups remained at approximately 0.5 standard deviations. The remaining group differences could indicate either that the matching procedure did not control adequately for administration selection effects or that there was an administration mode effect.

Way, Davis, and Fitzpatrick (2006) used matched sampling to compare mode effects and to use different score conversion tables for computer and paper groups when the equating relationships for the two groups be extremely discrepant. The authors conducted a simulation study to assess the sensitivity of the matched samples comparability analysis (MSCA) procedure given different magnitudes of mode effects. They found that the MSCA procedure did not always identify form differences of only 0.25 raw score points but that mode effects as large as 1 raw score point were always identified. Additional studies by the authors (McClarty, Lin, & Kong, 2009; Way, Lin, & Kong, 2008) also demonstrated the usefulness of matched sampling as a tool for evaluating mode effects.

Although matching appears to be a reasonable solution when equating with groups that differ substantially in performance, it may be difficult to include important variables in the matching process. Poor matching may result in bias, leading to equating results that are less accurate than when matching is not used. Because matching is currently used operationally to determine whether or not to use alternate score conversions for paper- or computer-based test takers (Way et al., 2006; Way et al., 2008), it is important to investigate the effects matching may have on results.

The purpose of this study was to investigate whether or not matched sampling can improve equating results. Specifically, the research questions this study seeks to address are:

1. Which matching techniques, if any, provide more accurate equating results?
2. Can matched samples equating reduce the extent to which equating assumptions are violated?

Method

Data Source

Data from the 2008 form of the Advanced Placement (AP) English Language mixed-format exam were included in this study. Although AP Exams were used in this study, several modifications were made to the scores and examinee groups. Therefore the focus of this study is on the application of equating methods generally, and not on the specific implications of the findings for AP Exams.

Multiple-choice (MC) items were operationally scored with a correction for guessing, also called formula scoring, to discourage examinees from randomly guessing. To eliminate missing data, examinees that responded to less than 80% of MC items were eliminated and a two-way imputation procedure (van Ginkel & van der Ark, 2005) was implemented to eliminate missing data for the remaining students. Finally, number correct scoring, where an examinee's score is the weighted sum of all correct responses to MC items and all score points obtained for free response (FR) items, was applied. The imputed data had a higher MC total score mean, were less variable, and were more negatively skewed (see Table 1). To avoid rounding noninteger scores, and so that the MC and FR sections would contribute to the composite score in the same ratios used operationally, integer weights of 2 for the MC section and 5 for the FR section were selected. The resulting composite scores ranged from 0 to 239.

The 2008 AP English Language Exam form was divided into two pseudo-forms which are hereafter referred to as Form A and Form B. MC items tied to a single passage were kept together and randomly assigned to a form. Some items were assigned to both forms to keep the forms of equal length. FR items were assigned to both forms. Common items included those used to link the operational 2008 form to previous forms, as these items are expected to be representative of the AP English Language Exam in terms of content and statistical properties. MC and FR items assigned to both pseudo-forms but not included in the operational 2008 common item set were considered non-common items for the purposes of equating.

Applying the CINEG design using pseudo-forms makes it possible to assess the statistical assumptions used with nonequivalent group equating methods because scores on both forms are known for both groups. The "new" Form A and "old" Form B groups are equivalent because the same examinees took both pseudo forms. Therefore the common item effect size (ES) is exactly zero. ES is calculated:

$$ES = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}}, \qquad (1)$$

where the subscript 1 refers to the new form group, and subscript 2 refers to the old form group. M stands for the common item mean, s^2 for the common item variance, and n for the sample size.

To study the impact of group differences on equating results, the ES was manipulated to achieve values of 0, 0.25, 0.5, and 0.75 by sampling examinees for Forms A and B using levels of parental education. Parental education was defined as the highest level of education obtained by either parent. Levels were coded into the following categories: 1) any education through high school diploma or equivalent, 2) trade school, some college, or associates degree, 3) Bachelors degree or some graduate/professional school, or 4) Graduate/Professional degree. To obtain group differences, more examinees were randomly sampled from within each of the lower categories of parental education and fewer examinees were randomly sampled from the higher categories of parental education for Form A than for Form B, resulting in a higher performing group for Form B. To equalize sampling error across conditions, 1500 students per form were sampled for each condition.

Equating Methods

CINEG equating methods included frequency estimation and chained equipercentile with cubic spline postsmoothing, and IRT true and observed score methods (Kolen & Brennan, 2004). Cubic spline postsmoothing was used in this study because it has been found to improve equating results and decrease the SE (Hanson, Zeng, Colton, 1994; Kolen, 1984). A smoothing value of S=.05 was used because it provided the smoothest equating relationship within a band of one SE around the unsmoothed equivalents. For the IRT observed score and frequency estimation methods, synthetic weights of 0.5 were used for both new and old form groups.

Item parameters were estimated with MULTILOG (Thissen, 1991) using the three-parameter logistic (3PL) model (Lord, 1980) for the MC items and the graded response model (GRM; Samejima, 1972) for the FR items. For IRT observed score equating, normal quadrature points and weights were used to approximate the ability distributions. Equating was conducted using *Equating Recipes*, a set of open source C functions (Brennan, Wang, Kim, & Seol, 2009). Items were calibrated separately by form and the Haebara (1980) scale transformation method was used to place parameter estimates on the same scale before equating.

Matching Methods

Matching old and new form groups has typically involved background variables, examinee scores on other related measures, or common item scores (e.g., Cook et al., 1988; Lawrence & Dorans, 1990; Schmitt et al., 1990; Way et al., 2006). When there are several variables used to match nonequivalent groups, Rosenbaum and Rubin (1983) propose using logistic regression to create a multivariate composite, which they call a propensity score. Once propensity scores are calculated, groups can be matched exactly (if possible), proportionally, based on stratification, or based on more complicated techniques (see Shadish, Cook, & Campbell, 2002, for a summary of techniques).

Operationally, the factors that differentiate group membership, called selection variables, are unknown. However, in this study, to simulate group differences, pseudo groups were created based on a known selection variable, parent education level. Equating results and assumptions were studied under the following four matching conditions:

No Matching (M_0): Equating was conducted using the non-equivalent pseudo groups created by sampling from parental education categories to achieve the desired ES. No matching was done.

Matching on Selection Variable (M_1): Non-equivalent pseudo groups were matched by randomly selecting the same number of examinees within each category of parental education. For example, if the sample sizes in pseudo group A were 200, 700, 500, and 500 and the sample sizes in pseudo group B were 500, 500, 700, and 300, then the largest number of examinees that could be matched would be 200, 500, 500, and 300.

Matching on Propensity Score Including Selection Variable (M_2): Pseudo groups were matched on a propensity score that included six background variables:

1. Gender
2. Fee indicator for low-income examinees
3. Region of the country (northeast, south, Midwest, and west) based on a student's reported state of residence and the US census definition of the geographical regions
4. Grade level
5. Parental education (the selection variable)
6. Ethnicity categorized into 5 highest-frequency categories (African American; Mexican or Mexican American; Puerto Rican, Latino, or other Hispanic; Asian; and White).

Matching was done using a SAS macro developed by Parsons (2001).

Matching on Propensity Score NOT Including Selection Variable (M_3): The same procedure was used as in M_2 except that parental education was excluded from the logistic regression equation used to calculate propensity scores.

Many testing programs use no matching (M_0) as their default equating procedure. If matching were to be used as part of an equating process, ideally, the matching would be done on the selection variable(s) (M_1). However, in practice, many background variables and scores may be available to use for matching, but it is not clear which variables are the selection variables. Therefore, if propensity scores are used to match examinee groups, there are two possible outcomes: either the propensity scores include the selection variable(s) (M_2), or they do not (M_3). In both cases, many variables may have been included in the model that were not selection variables. M_2 and M_3 are intended to represent the possible operational outcomes of using propensity score matching when the selection variable is unknown.

Evaluating Equating Results

Equating is known to work best when groups perform similarly (Kolen, 1990). Therefore, equating results based on groups with an ES of zero were considered the criterion equating relationship to which all other equating relationships were evaluated.

The SEs for the traditional criterion equating relationships were estimated using a bootstrap procedure with 1000 replications, and used to evaluate whether comparison group equating relationships were within sampling error of the criterion relationship. IRT SEs are not estimated by *Equating Recipes*, and were not included in this study. Instead, SEs for the traditional methods were used as an approximation. The SE was used for comparisons rather than the standard error of the equating difference (SEED) (von Davier, Holland, & Thayer, 2004) because the SEED is not part of the standard functions in *Equating Recipes*.

Root Expected Mean Squared Difference (REMSD) statistics (von Davier et al., 2004; Dorans & Holland, 2000;) were used to provide an overall indication of how equating relationships departed from the criterion relationship as ES increased.

$$REMSD = \frac{\sqrt{\sum_{\min(a)}^{\max(a)} v_{ac}[eq_{Bc}(a) - eq_{B0}(a)]^2}}{\sigma_0(B)}, \qquad (2)$$

where v_{ac} is the conditional proportion of Form A examinees at a particular Form A score for the comparison (c) equating relationship; $eq_{B0}(a)$ is Form B equivalent for the pseudo group with ES of zero; $eq_{Bc}(a)$ is the Form B equivalent for a comparison pseudo group with an ES greater than zero; and $\sigma_0(B)$ is the standard deviation of Form B scores for the ES of zero condition. The summation is taken over all Form A scores. REMSD was calculated by comparing the same equating method with the criterion group (ES = 0) as with the comparison group (ES = 0.25, 0.5, or 0.75).

AP scores are reported to students as AP grades that are integer scores ranging from 1 to 5. AP grades are intended to correspond roughly to college course grades where a 3 would indicate ‘C’ performance, a 4 would indicate ‘B’ performance, and so on. A majority of colleges require students to receive scores of at least 3 or 4 to receive credit or advanced placement. Therefore, classification consistency (CC) of AP grades provides an indication of the practical significance of differences in equating relationships. Classification consistency was calculated using the grades new group examinees received given the criterion equating relationship (ES = 0), and the AP grades they received with a comparison equating relationship (ES > 0).

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

where n_{aa} is the number of new group examinees that received a grade of *a* (where *a* is 1, 2, 3, 4 or 5) with both equating relationships, and *N* is the total number of examinees in the new group. Cut scores for Form B (the old form) were calculated so that the AP grade distribution was approximately equal on Form B and the operational 2008 AP English Language form.

Evaluation of Equating Assumptions

Direct evaluation of frequency estimation equating method assumptions involves comparing the conditional distributions of Form A scores given common item scores (*v*) in the new (1) and old (2) form groups [$f_1(A|V)$ v. $f_2(A|V)$], and making similar comparisons for Form B [$g_1(B|V)$ v. $g_2(B|V)$]. The conditional distributions are assumed to be the same in both groups ($f_1(A|V) = f_2(A|V)$ and $g_1(B|V) = g_2(B|V)$) (Kolen & Brennan, 2004). To quantify differences in $f(A|V)$ for the two groups, *v*+1 maximum differences were found between the Form A conditional cumulative frequency distributions for the two groups. The maximum differences were weighted by the sum of the Form A conditional frequencies across groups and then the *v*+1 weighted values were averaged. The same process was used to quantify differences in the Form B conditional cumulative frequency distributions for the two groups.

Chained equipercentile equating method assumptions can be directly evaluated by comparing the unsmoothed equipercentile equivalents for the link from *A* to *V* which are assumed to be the same in the old (2) and new (1) form groups ($e_{V1}(A)=e_{V2}(A)$), and by comparing the unsmoothed equipercentile equivlents from *V* to *B* which are assumed to be the same in both groups ($e_{B1}(V)=e_{B2}(V)$) (Holland, von Davier, Sinharay, & Han, 2006). Differences in the linking relationships for the two groups ($e_{V1}(A)$ v. $e_{V2}(A)$ and $e_{B1}(V)$ v. $e_{B2}(V)$) were quantified using REMSD statistics (von Davier et al., 2004; Dorans & Holland, 2000).

$$REMSD(e_V(A)) = \frac{\sqrt{\sum_{\min(A)}^{\max(A)} v_{A1}[eq_{V1}(A) - eq_{V2}(A)]^2}}{\sigma_2(V)},$$
$$REMSD(e_B(V)) = \frac{\sqrt{\sum_{\min(V)}^{\max(V)} v_{V1}[eq_{B1}(V) - eq_{B2}(V)]^2}}{\sigma_2(B)}. \quad (4)$$

Here v_{A1} is the conditional proportion of examinees at a particular Form A score for group 1; $eq_{V1}(A)$ is the common item equivalent for group 1; $eq_{V2}(A)$ is the common item equivalent for group 2; and $\sigma_2(V)$ is the standard deviation of common item scores for group 2. The summation for $REMSD(e_V(A))$ is taken over all Form A scores. For $REMSD(e_B(V))$, v_{V1} is the conditional proportion of examinees at a particular common item score (V) for group 1; $eq_{B1}(V)$ is the Form B equivalent for group 1; $eq_{B2}(V)$ is Form B equivalent for group 2; and σ_2(B) is the standard deviation of the Form B scores for group 2. The summation for $REMSD(e_B(V))$ is taken over all common item scores.

To investigate whether or not the IRT unidimensionality assumption (Lord, 1980) held for the data in this study, MC-FR correlations (uncorrected and disattenuated using coefficient alpha reliability estimates) and principal component analysis on polychoric correlations were computed. Parameter estimates for the MC section were estimated with and without the FR items and the results were compared. Likewise, FR item parameters were estimated with and without MC items to assess the stability of item parameter estimates.

Results

Matching

Table 2 provides phi coefficients for the pairs of background variables with values greater than 0.10. Phi was calculated for variables with two or more categories as the square root of the Pearson chi-square divided by the total number of observations (Conover, 1998, p. 234). Most of the phi coefficients were near zero. However, there was a moderate relationship between the fee indicator and parental education, between the fee indicator and ethnicity, between region and grade, between region and ethnicity, and between parental education and ethnicity.

Matching methods M_2 and M_3 involved the use of logistic regression to compute propensity scores for each examinee based on the matching variables. Table 3 provides the generalized (pseudo) r-squared values (Cox & Snell, 1989, p. 208-209) and p-values for the logistic regression coefficients for the intercept and the background variables entered into the model. As the ES increased, the r-squared values increased, indicating that background variables predicted group membership better for higher ESs. For M_2, this finding was expected because group differences in parental education increased as ES increased. For M_3, the r-squared values were substantially lower than those for M_2, which was expected because the selection variable was not included in the model. However, the r-squared values increased as ES increased for M_3 as well, which may have resulted from the correlation between parental education and the other background variables. At ES=0.75, the r-squared values for M_2 and M_3 were 0.626 and 0.145 respectively. For M_2, the intercept and parental education regression coefficients were always significantly different from zero ($p<0.05$). For M_3, the coefficients for all variables except high school grade level were significant.

The common item ES for the M_0, M_1, M_2, and M_3 matching methods are provided in Table 4. The ESs for the M_1 and M_2 matching methods were all relatively small, indicating that the groups were much more similar after matching than before matching. This finding was expected given that both methods matched on parental education. The M_3 ESs were only slightly lower than the M_0 ESs.

Comparison of Equating Relationships

Matched and unmatched equating results were compared for 12 combinations of ES and matching conditions. Plots comparing the old form equivalents for each matching method compared to the criterion equating relationship (ES=0) are provided in Figures 1 and 2. In Figure 1, plots are provided for the IRT true score equating method in the left column and for the IRT observed score equating method in the right column. The top plots are for an ES of 0.25, the middle for an ES of 0.5, and the bottom for an ES of 0.75. The equating relationships are plotted from approximately the first through the 99th percentiles. Equating relationships are provided in each plot for the unmatched condition (M_0), and the three matched conditions (M_1-M_3) using colored lines. The closer the equating relationships are to the vertical axis value of zero, the closer they are to the criterion equating relationship where ES=0. The two black lines represent plus and minus two SEs as calculated using 1000 bootstrap replications for the criterion chained equipercentile equating. Although these SEs may not be the same as the IRT bootstrap SEs, they provide an approximation that can be used to judge how different the equating relationships are. The same information is provided for the traditional equating methods in Figure 2. On the left are the results for frequency estimation, and on the right are the results for chained equipercentile.

Comparing the left three plots to the right three plots in Figure 1 indicates that both IRT methods provided nearly identical results. As ES increased, the deviation of M_0 and M_3 equating relationships from the criterion (vertical axis value of zero) increased. At ES=0.25, the M_0 and M_3 equating relationships differed from the criterion by nearly 10 score points at the low end of the scale; for ES=0.75, they differed by nearly 20 score points. For ES=0.25 and 0.50, M_1 produced the closest equating relationship to the criterion. M_1 results were within plus or minus two SEs from the criterion equating relationship for the majority of the score scale for all ES values, even at ES=0.75. M_2 produced equating results that were very close to the criterion for ES=0.75, but comparable to M_0 and M_3 for ES= 0.25 and 0.50. In fact, M_2 deviated the most from the criterion for ES= 0.25 and 0.50 at the high end of the scale. The sensitivity of the IRT equating results to the matching method appeared to increase as ES increased. For ES=0.75, the equating relationship for M_2 and M_0 differed by almost 15 points, or a half of a standard deviation, at the low end of the scale.

A comparison of the frequency estimation and chained equipercentile equating results in Figure 2 indicates that the equating results for the traditional methods were not nearly as similar as the equating results for the two IRT equating methods. At ES=0.25, M_0 and M_3 deviated from the criterion by more than two SEs for the majority of the score range for the frequency estimation method, and at some score points for the chained equipercentile method. The deviation of M_0 and M_3 from the criterion increased as the ES increases for both methods, but the frequency estimation method appeared more sensitive to the matching method. For ES=0.75, M_0 differed from the criterion by approximately 15 points (approximately 0.5 SD) at some scores for the frequency estimation method. M_1 and M_2 tended to stay within plus or minus two SEs even at ES=0.75.

M_0 results were compared to REMSD values for M_1-M_3. The REMSD values for all ESs, matching methods, equating methods, and exams are provided in Table 5. As expected, REMSD increased for the unmatched condition (M_0) as ES increased. M_1 REMSD values were consistently the smallest across equating and ES conditions. M_2 REMSD values were smaller than the M_0 and M_3 REMSD values for ES= 0.5 and 0.75.

IRT REMSD values tended to be lower for M_0 and M_3 than for the traditional methods, but they were not consistently lower for M_1 and M_2. The frequency estimation method had larger REMSD values than the other three methods for M_0 and M_3, but the M_1 and M_2 values were at least as small as the values for the other three methods. These results suggest that although the frequency estimation method is the most sensitive to large group differences, it benefits the most from matching.

Classification consistency values are provided in Table 6 for each comparison equating relationship where the classification for the ES=0 equating relationship was considered the criterion. Classification consistency did not decrease consistently from ES=0.5 to ES=0.75. M_0 and M_3 tended to have the lowest classification consistency values, especially for the frequency estimation method. M_1 and M_2 tended to provide an improvement over M_0, most noticeably for the frequency estimation method. M_3 tended to provide a small improvement in classification consistency values over M_0. Across all conditions, classification consistency ranged from 70.11-98.82%

Evaluation of Equating Assumptions

A weighted absolute maximum difference (WAMD) between the old form group and new form group cumulative frequency distributions was used to evaluate the frequency estimation assumptions. The WAMD between $f_1(A|V)$ and $f_2(A|V)$ were very similar to the WAMD between $g_1(B|V)$ and $g_2(B|V)$ so they were averaged and are provided in Table 7. As ES increased, the WAMD increased for M_0. In other words, as groups differences increased, the adequacy of frequency estimation assumptions decreased. However, the WAMD for M_1 and M_2 remained similar to the ES=0 value. WAMD for M_3 were slightly smaller than for M_0, but the M_1 and M_2 matching methods provided better results than M_3.

The WAMD values in Table 7 were compared to the REMSD values for the frequency estimation method in Table 5. As expected, higher WAMD (assumptions holding less well) in Table 7 corresponded to higher REMSD values (less accurate equating results) for the frequency estimation method.

Chained equipercentile equating assumptions were evaluated by calculating REMSD values comparing the linking relationship between the composite and common item scores for old and new form groups. $REMSD(e_V(A))$ and $REMSD(e_B(V))$ values were fairly similar so the values were averaged and included in the last column of Table 7. A comparison of the REMSD values for M_0 indicates that they increased consistently as ES increased. REMSD values for M_1 were smallest across all ESs. REMSD values for M_2 also provided an improvement over M_0

values at ES=0.5 and 0.75. The M_3 matching method did not appear to improve results over the M_0 method.

The REMSD values in Table 7 were compared to the REMSD values for the chained equipercentile method in Table 5. As expected, higher REMSD (assumptions holding less well) in Table 7 corresponded to higher REMSD values (less accurate equating results) for the chained equipercentile method.

Correlations between MC and FR items and principal components analysis were used to evaluate the IRT assumption of unidimensionality (not shown). ES did not have any predictable influence on the magnitude of the correlations. A comparison of the observed and disattenuated correlations for M_1-M_3 likewise did not reveal any systematic differences. The maximum disattenuated correlation found for English Language was 0.755, which indicates that the English Language Exam may not be unidimensional (see Powers, 2010).

Eigenvalues and scree plots were compared across ESs and matching methods and found to be nearly identical (not shown). A relationship between the dimensionality of the exam, as assessed by principal components analysis, and the ES or matching method used could not be found. MC and FR item parameter estimates were estimated together and separately, and the process was repeated for each of the matching conditions (M_0-M_3). No patterns were found to suggest that item parameter estimation was impacted by group differences or matching methods (see Powers, 2010).

Discussion

Consistent with other studies (Lawrence & Dorans, 1990; Livingston et al., 1990; Schmitt et al., 1990), the consistency of equating results was found to decrease as ES increased. Matched sampling has been considered in previous research as a possible way to improve the accuracy of equating results when groups differ substantially (e.g., Eignor et al., 1990; Lawrence & Dorans, 1990; Paek et al., 2006). This study compared the equating accuracy of four matching methods and evaluated the impact of matching on the degree to which equating assumptions were met.

Based on the results of this study, it appears that IRT true and observed score equating results are nearly identical even when the ES is very large. Lord and Wingersky (1984) also reported that the IRT true and observed score equating methods provided very similar results. One possible explanation for this finding is that equating accuracy appears to be related to the degree to which equating assumptions are violated, and both IRT methods share the same unidimensionality assumption.

The IRT and chained equating results appear to be less sensitive to group differences than the frequency estimation method. However, the accuracy of the frequency estimation method appears to be greatly improved by matching on the selection variable. As mentioned previously, findings about the sensitivity of equating methods to sampling method have been mixed (Cook et al., 1988, 1990; Eignor et al., 1990; Lawrence & Dorans, 1990; Livingston et al., 1990; Schmitt et al., 1990; Wright & Dorans, 1993). However, previous studies involved fairly modest group differences, which may have obscured the relationship between ES and equating accuracy.

For frequency estimation and chained equipercentile equating methods, it appears that matching on the selection variable when group differences are large improves the degree to which equating assumptions are met, thereby increasing the accuracy of equating results. For IRT equating methods, although a direct link between group differences and equating assumptions could not be established, matching on the selection variable still improved equating accuracy.

Whereas other studies have used matched samples equating (e.g., Eignor et al., 1990; Lawrence & Dorans, 1990; Paek et al., 2006), and even used matched sampling operationally (Way et al., 2006; Way et al., 2008), the methods used in this study made it possible to match using the true selection variable. Because the selection variable was known, it was possible investigate how different matching methods would impact the degree to which equating assumptions held and the accuracy of equating results. Although matching on only the selection variable (M_1) was considered the best-case-scenario for matched samples equating, an important finding was that matching on a propensity score where the selection variable was included in the logistic regression (M_2) provided reasonably comparable results.

The finding that matching on a propensity score that did not include the selection variable (M_3) did not provide improvement in the accuracy of equating results is also important. M_3 included variables that had a fairly low correlation with the selection variable. It is not surprising therefore that matching was not successful in eliminating group differences, improving the degree to which equating assumptions were met, or increasing the accuracy of equating results. Because the relationship between M_3 and parental education was so modest, M_3 was almost a worst-case-scenario for matching with variables that do not include the selection variable.

Researchers have used propensity score matching with prior years' test scores (McClarty et al., 2009; Way et al., 2006; Way et al., 2008). Although the analyses used in this study did not address the efficacy of using prior test scores as matching variables, the use of prior test scores as matching variables seems like a promising method that may correlate much higher with selection variables than the many of the variables included in this study.

Shadish et al. (2002) warn that inappropriate matching can increase bias. Although there was no conclusive evidence that M_3 had detrimental effects on equating results, it may have introduced some equating bias. It appears that propensity score matching can be beneficial, especially with very extreme group differences, when the selection variable or set of variables is included in the logistic regression. However, if the propensity score does not happen to include the unknown selection variable(s), or at least variables that are highly correlated with the selection variable, the matching process may introduce more equating error than an unmatched samples equating. Variables that are very highly correlated with the selection variable, for example, previous years' test scores on related measures, may provide improved results, but additional research is necessary to confirm this hypothesis.

One important component of this study was the use of real data rather than simulated data to assess the impact of group differences on equating results. However, the use of real data limited the number of replications that could be used to draw conclusions. Only one sample was selected for each ES. In addition, a number of modifications were made to the scores and groups that may limit the generalizability of the results.

Also, the initial sampling on a single selection variable (parental education) to obtain "old" and "new" form groups that differ by target ESs, created nonequivalent groups that would be unlikely to have occurred in practice. In practice it is likely that there are several variables that explain group differences, rather than the single selection variable used in this study. Moreover, many of the "true" selection variables may be difficult or impossible to measure. Even if the selection variables could be measured, there is always limited administration time with which to measure the variables.

If the magnitude of the common item ES is responsible for assumption violations and inaccurate equating results, it might seem reasonable to match groups based on common item scores only. A limitation of this study is that a common item matching method was not considered. However, researchers at ETS have used common item matching with little success

(Kolen, 1990). The clear benefit of a common item matching method is that common item scores are available for all examinees. No additional administration time is needed.

This study was also limited in scope in that the impact of common item composition was not considered. Although the AP Exams are mixed-format, operationally, the common item set only includes MC items. Likewise, only MC common items were used in this study. Representative common item sets can produce less biased equating results (Kolen, 1990). The frequency estimation and chained equipercentile assumptions involve the common items. A change in their composition would likely change both the adequacy of the equating assumptions, and the accuracy of equating results. However, the inclusion of FR items in the common item set can also be problematic because of security concerns and rater drift.

Issues of composite-to-common item ratios were also not considered. Also there were a limited number of equating methods considered. Other methods that might be investigated include linear equating, multidimensional IRT models, and IRT models that are designed to handle testlets. Finally, only one smoothing value was considered in this study for postsmoothed traditional equating relationships. The IRT equating results were much smoother than the traditional equating results, indicating that a higher smoothing value may have made equating results more comparable. The criteria used to select *S*-values in this study may have been too stringent. Conclusions may be impacted by the degree of smoothing chosen.

Additional research is recommended to include more diversity of datasets with large ES ranges, additional samples at each level of ES and matching method, more realistic selection variable(s), the use of matching variables that are more highly correlated with the selection variable, consideration of the representativeness of common item sets in mixed-format exams, use of additional equating methods, and additional investigation of conditions that increase CINEG equating assumption violations.

References

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes* (CASMA Monograph Number 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Conover, W. J. (1999). *Practical nonparametric statistics (3rd ed.).* Wiley.

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

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (ETS Research Report RR-88-52). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1990). *Equating achievement tests using samples matched on ability* (ETS Research Report RR-90-10). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1998). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31-45.

Cox, D. R. & Snell, E. J. (1989). *The analysis of binary data (2nd ed.).* London: Chapman and Hall.

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

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3*, 37-52.

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

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: ACT.

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design.* (ETS Research Report RR-06-17). Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25-44.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*, 97-104.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices. (2nd ed.)* New York: Springer-Verlag.

Lawrence, I. M., & Dorans, N. J. (1990). Effects on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*, 19-36.

Livingston, S. A., Dorans, N. J., & Wrights, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73-95.

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

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

McClarty, K. L., Lin, C.-H., & Kong, J. (2009). *How many students do you really need? The effect of sample size on matched samples comparability analyses.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Paek, I., Liu, J., & Oh, H.-J. (2006). *Investigation of propensity score matching on linear/nonlinear equating method for the PSAT/NMSQT* (ETS Statistical Report SR-2006-55). Princeton, NJ: Educational Testing Service.

Parsons, L. S. (2001). *Reducing bias in a propensity score matched-pair sample using greedy matching techniques.* Paper 214-26. Retrieved October 5, 2009 from http://www2.sas.com/proceedings/sugi26/p214-26.pdf

Powers, S. (2010). *Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions.* Unpublished doctoral dissertation, University of Iowa.

Powers, S., & Kolen, M. J. (2011). *Evaluating equating accuracy and assumptions for groups that differ in performance.* In M. J. Kolen, & W. Lee (Ed.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 1)* (CASMA Monograph No. 2.1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*, No. 18. Retrieved July 1 2010, from http://www.psychometrika.org/journal/online/MN18.pdf

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3*, 53-71.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT pre-equating* (ETS Research Report RR-86-49). Princeton, NJ: Educational Testing Service.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International, Inc.

van Ginkel, J. R., & van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, 29, 152-153.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, *41*, 15-32.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills* (PEM Research Report 06-01). Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Way, W. D., Lin, C.-H., & Kong, J. (2008). *Maintaining score equivalence as tests transition online: Issues, approaches and trends.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS Research Report RR-93-4). Princeton, NJ: Educational Testing Service.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.

Table 1

Unweighted MC Moments in Original and Imputed Data

Data	N	M	SD	Skew	Kurt
Original	301,095	29.37	12.05	-0.25	-0.58
Imputed	247,197	38.12	9.48	-0.74	0.12

Table 2

Phi Coefficients for Examinee Background Variables

N	214,049
Fee & Parent Ed	0.34
Fee & Ethnicity	0.36
Region & Grade	0.20
Region & Ethnicity	0.29
Parent ED & Ethnicity	0.34

Note. Values of phi for all other pairs of variables were less than 0.10.

Table 3

Significance of Variables in Logistic Regression Equation

			Probability > Chi Square						
ES	Method	R^2	Intercept	Gender	Ethnicity	Grade	Fee	Region	Parent ED
0.25	M_2	0.21	<0.01	0.92	0.17	0.84	<0.01	0.67	<0.01
	M_3	0.02	<0.01	<0.01	<0.01	0.62	<0.01	<0.01	--
0.50	M_2	0.57	<0.01	0.68	0.11	0.25	0.20	0.53	<0.01
	M_3	0.06	<0.01	<0.01	<0.01	0.30	<0.01	<0.01	--
0.75	M_2	0.63	<0.01	0.01	0.26	<0.01	0.31	0.32	<0.01
	M_3	0.14	<0.01	<0.01	<0.01	0.74	<0.01	<0.01	--

Table 4

Changes in ES for Matched and Unmatched Groups

ES	M_0	M_1	M_2	M_3
0.25	0.24	-0.02	0.04	0.17
0.50	0.44	-0.04	0.03	0.38
0.75	0.72	-0.03	0.00	0.54

Table 5

REMSD Values

		Equating Method			
ES	Method	FE	CE	IRT True	IRT Obs
0.25	M_0	0.12	0.09	0.09	0.09
	M_1	0.04	0.05	0.05	0.04
	M_2	0.08	0.10	0.12	0.11
	M_3	0.11	0.09	0.08	0.08
0.5	M_0	0.23	0.13	0.15	0.15
	M_1	0.05	0.06	0.06	0.05
	M_2	0.08	0.09	0.11	0.09
	M_3	0.22	0.15	0.16	0.15
0.75	M_0	0.33	0.16	0.22	0.20
	M_1	0.04	0.06	0.08	0.08
	M_2	0.04	0.04	0.03	0.03
	M_3	0.29	0.18	0.17	0.17

Notes. FE = postsmoothed frequency estimation.
CE = postsmoothed chained equipercentile.

Table 6

Classification Consistency with ES=0 as the Criterion

		Equating Method			
ES	Method	FE	CE	IRT True	IRT Obs
0.25	M_0	89.48	91.29	95.44	94.12
	M_1	96.98	95.48	95.17	97.07
	M_2	93.58	94.15	90.14	91.97
	M_3	89.21	91.7	93.13	91.81
0.50	M_0	75.18	84.37	86.68	85.36
	M_1	96.31	94.61	95.21	96.53
	M_2	95.87	93.09	91.32	93.63
	M_3	76.39	86.87	92.84	91.52
0.75	M_0	70.11	86.97	89.66	91.59
	M_1	95.69	94.54	95.44	94.12
	M_2	98.06	97.49	97.5	98.82
	M_3	71.64	86.63	92.15	93.47

Note. FE = frequency estimation. CE = chained equipercentile.

Table 7

Evaluating the FE and CE Equating Assumptions

ES	Method	FE: Ave. WAMD	CE: Ave. REMSD
0	M_0	16.14	0.04
0.25	M_0	26.70	0.08
	M_1	17.73	0.04
	M_2	20.83	0.10
	M_3	25.29	0.08
0.5	M_0	42.06	0.15
	M_1	19.49	0.05
	M_2	15.04	0.09
	M_3	38.93	0.14
0.75	M_0	66.39	0.16
	M_1	20.58	0.05
	M_2	19.12	0.05
	M_3	57.54	0.16

Note. FE = frequency estimation. CE = chained equipercentile.
WAMD = weighted absolute maximum difference.
REMSD = root expected mean squared difference.

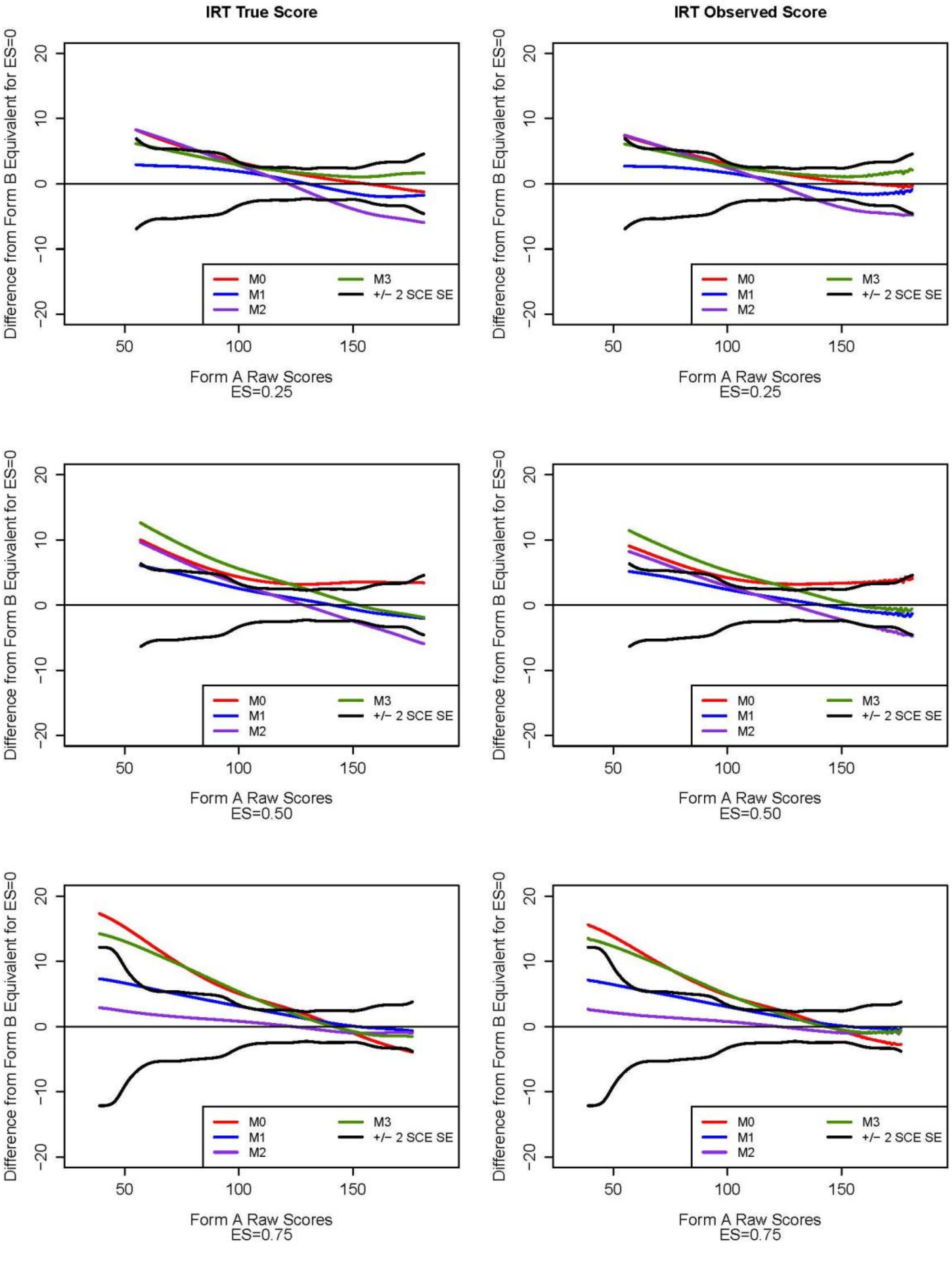


Figure 1. Comparison of matching methods across ES levels for IRT equating methods.

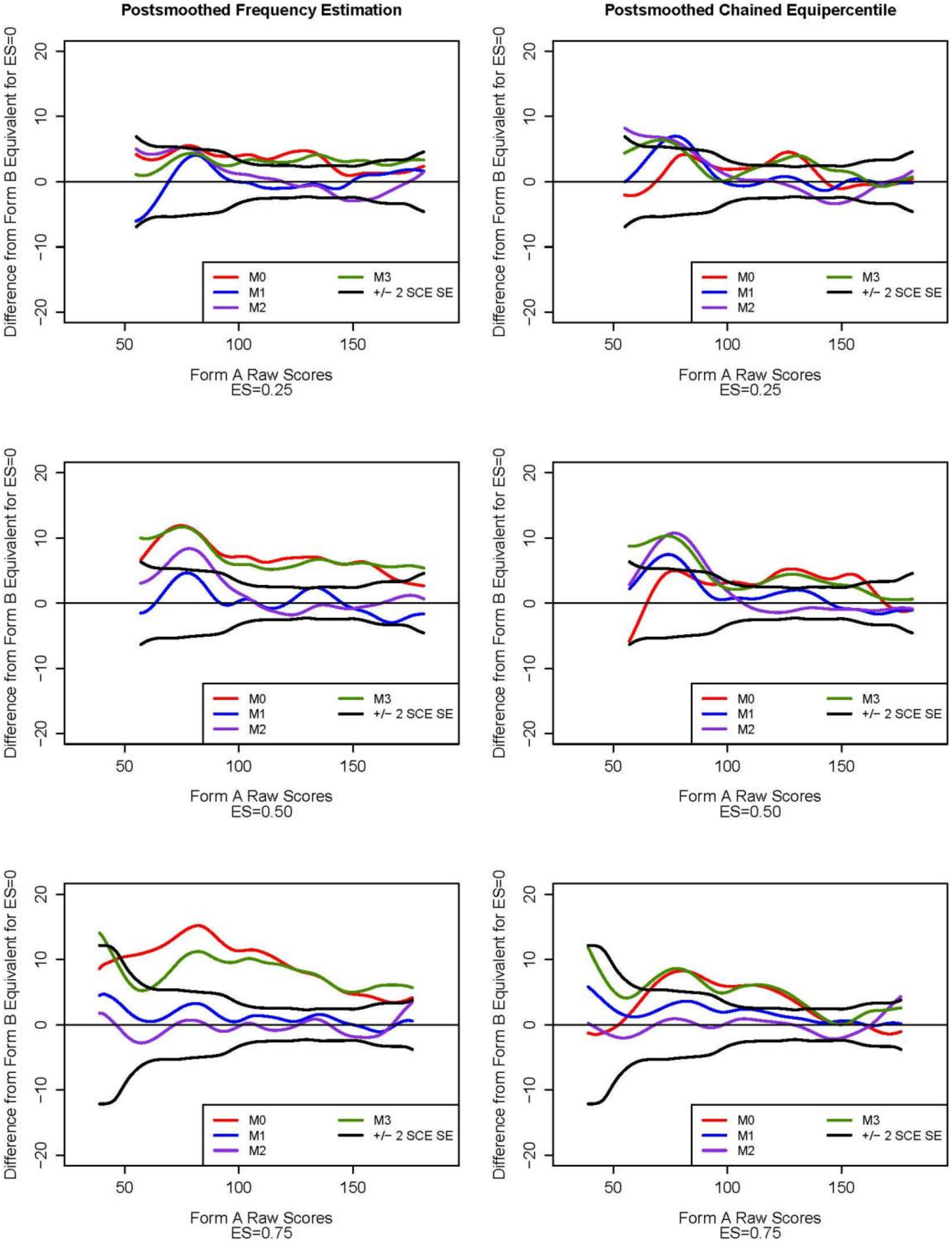


Figure 2. Comparison of matching methods across ES levels for traditional equating methods.

Chapter 5: Observed Score Equating for Mixed-Format Tests Using a Simple-Structure Multidimensional IRT Framework

Won-Chan Lee
The University of Iowa, Iowa City, IA
Bradley G. Brossman
American Board of Internal Medicine, Philadelphia, PA

Abstract

For many large-testing programs, equating is an integral part of the test development process. Therefore, it is imperative to ensure that equating results are as accurate as possible. If a unidimensional IRT equating procedure is applied to tests that have a multidimensional data structure, the resulting equating relationships will likely be inaccurate. When data are multidimensional, a multidimensional IRT approach would likely yield more accurate results. This paper proposes an IRT observed-score equating procedure under the simple-structure multidimensional IRT (SS-MIRT) framework that can be used to equate tests that consist of a pre-specified set of item clusters, each of which is associated with a single proficiency. In this chapter, the SS-MIRT equating procedure is applied to a mixed-format test, in which each of the multiple-choice and free-response item types is associated with separate, yet correlated, proficiencies. Results for both a real data analysis and a simulation study reveal that, when data are multidimensional, the SS-MIRT procedure produces adequate equating results and outperforms the traditional unidimensional IRT procedure.

Observed Score Equating for Mixed-Format Tests Using a Simple-Structure Multidimensional IRT Framework

A growing number of tests in large-scale testing programs are mixed-format tests, in which there exists a mix of multiple-choice (MC) and free-response (FR) items. For these tests, composite scores are of interest for the purposes of score reporting and psychometric analyses (Kolen, 2006). Although a mixed-format test is often intended to measure a single general proficiency (or construct), it is very likely that different item types measure somewhat different, yet highly correlated, proficiencies (Cao, 2008; Li, Lissitz, & Yang, 1999; Sykes, Hou, Hanson, & Wang, 2002; Tate, 2002; Yao & Boughton, 2009). When a mixed-format test is viewed as measuring a composite of two distinct (yet correlated) proficiencies (i.e., MC items measure only an MC-related proficiency and FR items measure only an FR-related proficiency), the structure of the data conforms to what is referred to here as a simple-structure multidimensional model. Kolen, Wang, and Lee (2012) used a simple-structure multidimensional model in the context of estimating conditional standard errors of measurement for composite scores. The primary purpose of the present paper is to present an observed-score equating procedure using a simple-structure multidimensional item response theory (SS-MIRT) model under a random groups design.

A great deal of research has been conducted in multidimensional item response theory (MIRT; Reckase, 2009) and its application to scale linking (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima, Davey, & Lee, 2000; Yon, 2006). However, little research exists in the literature that deals with test equating using MIRT models. One of the first attempts to equating in an MIRT framework is the study by Brossman and Lee (2012) who proposed observed- and true-score equating procedures in conjunction with a multidimensional two-parameter logistic model for a random groups design. The framework used in Brossman and Lee (2012) might be viewed as a complex MIRT model in that each item in a test was allowed to measure more than one proficiency as specified in the model.

As discussed in detail later, the SS-MIRT approach proposed in the present paper is different from the complex MIRT framework in many aspects. First, each item in the test under the SS-MIRT framework is associated with only one proficiency, which makes the calibration process much easier by fitting a unidimensional IRT model to a set of items grouped together according to a pre-specified dimensional structure. Moreover, the correlations among proficiencies are explicitly considered and estimated, which is not necessarily the case for the complex MIRT framework. While the complex MIRT framework requires a scale linking procedure (e.g., Thompson, Nering, & Davey, 1997) even for a random groups design to resolve rotational indeterminacy, the SS-MIRT framework does not. In addition, the SS-MIRT framework provides a straightforward interpretation of dimensionality and an effective treatment of weights associated with different dimensions. The SS-MIRT framework for equating is considered appropriate when items are grouped together according to a pre-specified dimensional structure such as content domains or item formats.

The purpose of the present study is to:

1. Present an observed-score equating procedure based on the SS-MIRT framework under a random groups design;
2. Compare results for the SS-MIRT equating procedure to results for other alternative equating methods using real data sets; and
3. Evaluate the performance of the SS-MIRT equating procedure using a simulation study.

SS-MIRT Equating Procedure

Although the SS-MIRT equating procedure can be used for any number of clusters of items, in this section it is applied to mixed-format tests consisting of MC and FR items. Let Form X and Form Y indicate the new and old forms, respectively. A unidimensional IRT model is assumed for each item type. Although any combination of unidimensional models could be used, in this paper, the three-parameter logistic model (3PL; Lord, 1980) is assumed for MC items and the graded response model (Samejima, 1997) is assumed for FR items.

The observed-score equating procedure using the SS-MIRT framework is based on the following assumptions: (a) Each item in a mixed-format test measures a proficiency corresponding to a specific item type, and proficiencies associated with different item types are allowed to be correlated; (b) a subset of items with the same item type can be modeled adequately by a unidimensional IRT model; and (c) examinees in the old and new form groups are randomly equivalent in terms of both the MC-related and FR-related proficiencies. The following steps are involved in the SS-MIRT observed-score equating procedure:

1. Calibrate MC and FR sections separately for each form.
2. Compute the conditional observed-score distribution for each item type for each form.

3. Compute the conditional total-score distribution over the two item types for each form.
4. Estimate a bivariate (MC and FR) proficiency distribution.
5. Compute the marginal total-score distribution using the estimated bivariate proficiency distribution for each form.
6. Conduct the equipercentile equating based on the two marginal total-score distributions for Form X and Form Y.

The above steps are very similar to the equating processes of the typical unidimensional IRT observed-score equating with the exception that a bivariate proficiency distribution and composite total-score distribution are involved. Each of the steps noted above is discussed in detail below.

Calibration

The MC-item section and FR-item section are calibrated separately for each of Form X and Form Y, which leads to four separate calibrations. Because the two groups are randomly equivalent, no scale linking process is required if the scale indeterminacy is resolved by specifying the proficiency distributions to be equivalent across calibrations. That is, parameter estimates for the MC items in Form X and Form Y are on the same MC proficiency scale; likewise, the FR item parameter estimates on the two forms are on the same FR proficiency scale. Also, the correlation between the MC and FR proficiencies is assumed to be equal for the two groups under the random groups design.

Conditional Observed-Score Distributions

Based on the estimated item parameters, the conditional observed-score distribution at each pair of θ_1 (MC proficiency) and θ_2 (FR proficiency) is obtained for each item section separately for each form. Let $f_1(x_1 \mid \theta_1) = \Pr(X_1 = x_1 \mid \theta_1)$ and $f_2(x_2 \mid \theta_2) = \Pr(X_2 = x_2 \mid \theta_2)$ denote the conditional observed-score distributions for MC and FR scores, respectively. These conditional distributions can be produced using a recursive algorithm presented by Lord and Wingersky (1984) or an extended version of the Lord-Wingersky algorithm discussed in Hanson (1994) and Thissen, Pommerich, Billeaud, and Williams (1995). Let $X = w_1 X_1 + w_2 X_2$ define the total observed score, which is a sum of weighted MC and FR scores[1]. Assuming conditional independence, the conditional total-score distribution can then be computed as

$$f(x \mid \theta_1, \theta_2) = \Pr(X = x \mid \theta_1, \theta_2) = \sum_{X = w_1 X_1 + w_2 X_2} f_1(x_1 \mid \theta_1) f_2(x_2 \mid \theta_2), \tag{1}$$

where the summation is taken over all possible pairs of $w_1 x_1$ and $w_2 x_2$ scores that lead to a particular total score *x*.

Marginal Observed-Score Distributions

A marginal observed-score distribution is obtained by aggregating conditional total-score distributions over an entire bivariate theta distribution, $g(\theta_1, \theta_2)$, which is given by

$$f(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x \mid \theta_1, \theta_2) g(\theta_1, \theta_2) d\theta_1 d\theta_2 . \tag{2}$$

If a set of quadrature points is used to represent the proficiency distribution, integration is replaced by summation as

$$f(x) = \sum_{\theta_1} \sum_{\theta_2} f(x \mid \theta_1, \theta_2) q(\theta_1, \theta_2), \tag{3}$$

where $q(\theta_1, \theta_2)$ is the density associated with a particular pair of θ_1 and θ_2, and $\sum q(\theta_1, \theta_2) = 1$. The marginal observed-score distribution is computed for each of Form X and Form Y (i.e., $f(x)$ and $f(y)$) based on item parameter estimates and a bivariate theta distribution.

The joint bivariate proficiency distribution, $g(\theta_1, \theta_2)$, can be estimated using Mislevy's (1984) procedure. Among several alternative latent distributions that Mislevy considered, a bivariate standard normal distribution is assumed in this paper to estimate the means and variances of θ_1 and θ_2 and the correlation between them. However, since the estimated mean and variance typical are close to zero and one, respectively, only the correlation is estimated here. Note that only one bivariate theta distribution is needed for equating because the two groups are assumed to be from the same population.

Equipercentile Equating

If the IRT models fit the data adequately, the fitted observed-score distributions, $f(x)$ and $f(y)$, will be smooth and will approximate the actual frequency distributions well.

[1] Integer weights are assumed here for programmatic reasons, although there is nothing in theory that requires it.

Traditional equipercentile is then conducted to determine the equating relationship between Form X and Form Y using the two fitted observed-score distributions.

An Illustrative Example

Data Source and Analysis

The SS-MIRT procedure was applied to equate two forms of the Advanced Placement (AP) World History test. Note that this illustrative example based on the data from this particular test was different in many aspects from the operational equating for the test—i.e., (a) number-correct scoring was used instead of operational formula scoring; (b) the score scale used in this study was not the one that is used operationally; and most importantly, (c) a random groups design was used as opposed to the operational common-item nonequivalent groups design. Thus, this example was intended for illustrative purposes only.

Each of the new and old forms had 70 MC items scored 0 or 1, and 3 FR items scored 0-9. The total score was a weighted sum of the MC and FR sections with $w_1 = 1$ and $w_2 = 2$, which resulted in total scores ranging from 0 to 124. The total scores were transformed to normalized scale scores ranging from 0 to 70. Table 1 provides the raw-to-scale score conversion table for the old form. The sample size was about 6,000 for each form. Equating results were obtained for both the raw and scale scores. The 3PL model was used for the MC items and graded response model was used for the FR items. Separate calibration was conducted for each form and each section using PARSCALE (Muraki & Bock, 2003). Results for the SS-MIRT procedure were compared to results for the unidimensional IRT (UIRT) observed-score equating procedure and the equipercentile method in conjunction with log-linear presmoothing with a parameter of 6 (i.e., *C* = 6). The UIRT and equipercentile equating with presmoothing were conducted using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009).

The observed correlation between the MC and FR scores was .76 and .77 for the new and old forms, respectively. The correlation corrected for unreliability using alpha as a reliability estimate was .91 and .94 for the new and old forms, respectively.

Results of the Illustrative Example

The extent to which the SS-MIRT and UIRT models fit the data adequately was evaluated by comparing the fitted marginal observed-score distributions based on the fitted models and the actual frequency distributions. Table 2 presents the first four moments of the actual and fitted observed-score distributions for the new and old forms. It appears that the moments of the fitted distribution for the two models are remarkably similar and tend to be close to the moments of the actual distributions. Figure 1 displays the actual and fitted distributions along the score scale. Notice that, in addition to the results for the two IRT models, a fitted distribution based on the loglinear presmoothing using $C = 6$ is shown in the figure. The fitted distributions for the two IRT models are very similar to each other and follow the actual frequency distributions well, in general. The log-linear fitted distributions are closer to the actual frequency distributions than the IRT results, but they seem to show a tendency for overfitting (especially for the new form) due to the use of a relatively large parameter value of 6 (i.e., the first six moments are preserved).

The correlation between the latent MC and FR proficiencies was estimated using the Mislevy (1984) procedure. The estimated correlations were .91 and .93, respectively, for the new and old forms. Note that these estimated latent correlations were very similar to the disattenuated correlation values of .91 and .94. Since the new and old form samples are assumed to be from the same population under the random groups equating design, a bivariate standard normal distribution with a correlation of .92 (i.e., the average correlation) was used to define the population for equating.

Equating results for raw scores are displayed in Figure 2 in terms of differences in the results between each equating method and the identity equating (i.e., no equating). The results for all three equating methods show a similar curvilinear equating relationship. It is interesting to note that the results for the SS-MIRT procedure are more similar to the results for the equipercentile method than to the results for the UIRT procedure. One plausible explanation for this observation might be that when the data are not strictly unidimensional, the UIRT procedure violates the assumption of unidimensionality. On the other hand, the SS-MIRT procedure takes into account the multidimensionality resulting from having a mix of item formats, and the equipercentile method is silent about the data dimensionality and thus is less affected by the dimensional structure of the test (Brossman & Lee, 2012).

In order to examine the magnitude of the differences in the equating results for the UIRT and SS-MIRT procedures, the differences in the equated scores are plotted against a ±0.5 band. The value of .5 for the band was selected to be consistent with the suggestion for Differences That Matter (DTM) (Dorans & Feigenbaum, 1994). Figures 3 and 4 display results for the raw scores and unrounded scale scores, respectively. For both equated raw and scale scores, the differences in the results for the two IRT equating methods went outside of the band across a wide range of the score scales. This suggests that the two IRT equating methods can produce somewhat different equating results even when the data demonstrate only weak multidimensionality (i.e., a latent correlation of .92).

The first four moments for equated scores are summarized in Table 3. Among the three equating methods, the equipercentile method produced equating results with moments that are closest to the old-form moments. Between the two IRT methods, however, the SS-MIRT procedure slightly outperformed the UIRT procedure in terms of the moments for the equated scores being closer to the moments for the old form for all three types of score scales.

A Simulation Study

A simulation study was conducted to evaluate the performance of the SS-MIRT procedure and incorporated varying levels of the correlation between the MC and FR proficiencies as a factor under consideration. This simulation study was intended to provide information regarding when the SS-MIRT and UIRT equating procedures produce adequate equating results if the test data conform to a simple-structure multidimensional model. Note that the traditional equipercentile equating method was not considered in the simulation study because the focus was on the comparison of the two IRT procedures and the simulation process including the determination of true equating relationships was entirely based on IRT.

Parameters

The estimated item parameters for the two forms of the World History test that were used in the previous section for an illustrative example were used here again as generating or “true” item parameters. The weights of 1 and 2 were used for the MC and FR sections, which led to a total-score range of 0-124. The same score scale ranging from 0 to 70 was used. The 3PL model was assumed for the MC items, and the graded response model was assumed for the FR items. The random groups design was used, in which two equating samples were assumed to come from the same population of a bivariate standard normal distribution, $N(0,0,1,1,\rho)$ with varying levels of ρ. A set of 1,681 pairs (41 x 41) of bivariate quadrature points and weights were used. The range of theta values for each proficiency was -4 to 4.

True equating relationships were established by conducting the SS-MIRT procedure based on the generating item parameters and the population proficiency distribution. True equating relationships were obtained for raw scores, unrounded scale scores, and rounded scale scores, and were compared to the sample-based estimated equating relationships. Note that different true equating relationships were obtained for the three levels of proficiency correlation.

Simulation Procedure

Three levels of correlation between the MC and FR proficiencies were considered: .95, .8, and .5. The sample size was fixed at 3,000 per form. The following simulation process was used:

1. Draw three thousand pairs of true theta values (*N* = 3,000) from $N(0,0,1,1,\rho)$ for each of the new and old form groups.
2. Generate item-response strings for each examinee for each form based on the generating item parameters and true theta values using the 3PL model for the MC items and the graded response model for the FR items.
3. Conduct the SS-MIRT and UIRT equating procedures based on the sample data to obtain estimated equating relationships.
4. Repeat the above steps 100 times (*R* = 100).
5. Repeat the above steps for three different levels of the proficiency correlation.

It should be noted that when the SS-MIRT procedure was applied to the sample data, the true correlation value was used to define the equating population, rather than an estimate based on the Mislevy's procedure. This was done to simplify the simulation process, and a preliminary analysis had shown that a small variation in the estimated correlation values had only minimal impacts on the equating results. For the UIRT procedure, the standard normal distribution with 41 quadrature points and weights was used as the equating population for all replications.

Evaluation Criteria

The estimated equating results for each of the two IRT equating methods over 100 replications were compared to the true equating relationships using three summary statistics: mean squared error (MSE), squared bias (SB), and variance (VAR). These three summary statistics were computed for each raw-score point first and then aggregated across all raw-score points using the new-form fitted frequency distribution as weights. The SB at raw-score point *x* is given by

$$SB(x)=\left[\left(\frac{1}{R}\sum_{r=1}^{R}\hat{e}_{xr}\right)-e_x\right]^2, \qquad (4)$$

where e_x is the true equated score at score *x*, and $\hat{e}_{xr}$ is an estimated equated score at score *x* on replication *r*. The VAR at raw-score point *x* is given by

$$VAR(x) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{e}_{xr} - \left(\frac{1}{R}\sum_{r=1}^{R}\hat{e}_{xr}\right)\right]^2 . \quad (5)$$

The MSE at score *x* is

$$MSE(x) = SB(x) + VAR(x) . \quad (6)$$

The square root of Equation (6) is the root mean squared error, $RMSE(x)$. The marginal SB across all raw-score points is computed by

$$SB = \sum_{x=0}^{\max(x)} h(x)SB(x) , \quad (7)$$

where $h(x)$ is the relative frequency distribution (which sums to one) for the new-form scores based on the assumed IRT model. The marginal VAR and the marginal MSE were obtained in a similar manner. These summary statistics were computed for equated raw scores, unrounded scale scores, and rounded scale scores for each equating method.

The $RMSE(x)$ statistic was compared to a DTM value of .5 to determine whether the overall amount of error indexed by $RMSE(x)$ across the range of the score scale is practically unacceptable. Both $RMSE(x)$ and the .5 DTM value were standardized using the standard deviation of true equated scores for each of the three score scales.

Results of the Simulation Study

Table 4 presents three marginal summary statistics for the two equating methods at each combination of the three levels of correlation and the three score scales. As anticipated, across all conditions, the results for the SS-MIRT procedure show substantially smaller error than the results for the UIRT procedure for all three summary statistics. The differences between the results for the two equating methods become smaller when the rounded scale scores are considered. Note that, for both equating methods, the VAR tends to contribute more to the MSE, relative to the SB. The results for the VAR statistics show that the variance for the UIRT procedure is substantially larger than the variance for the SS-MIRT procedure. This might be due, in part, to the number of quadrature points that was used for equating (i.e., 1,691 for SS-MIRT vs. 41 for UIRT), but primarily, it is the consequence of the model misfit of the UIRT procedure. Focusing on the SB, the bias tends to decrease as the correlation increases for the

UIRT procedure; while the opposite is true for the SS-MIRT procedure—namely, the SS-MIRT tends to have smaller bias when the correlation is lower. The particular pattern of the relationship between the bias and correlation levels for the SS-MIRT procedure deserves further investigation. It is clear that the difference between the UIRT and SS-MIRT procedures in terms of bias tends to increase as the correlation decreases for all three types of score scales. The MSE and VAR statistics do not seem to have any clear pattern in relation to the correlation levels.

The standardized $RMSE(x)$ values are plotted against the standardized DTM in Figures 6 through 8 for the correlation conditions of .5, .8, and .95, respectively. Across all three levels of the correlation, the $RMSE(x)$ values for the SS-MIRT procedure are lower than those for the UIRT procedure, and fall below the standardized DTM throughout most score levels for both raw and unrounded scale scores, except at the lower end of the score scale when $\rho = .8$. By contrast, the $RMSE(x)$ for the UIRT procedure in terms of the raw scores tends to exceed the standardized DTM at many score points. However, the results of the unrounded scale scores for the UIRT procedure tend to be lower than the DTM across a wide range of the score scale when the correlation is .8 or .95, indicating acceptable equating results.

Conclusions and Discussion

When parallel forms of a test are created and administered, equating becomes an integral part of the testing program. It is imperative to ensure that equating results are as accurate as possible. The accuracy of equating can be jeopardized if, for example, the test administration conditions are not standardized, a security breach occurs, the sampling process is not followed properly, the statistical assumptions of an equating method are violated, etc. The traditional unidimensional IRT equating procedure relies on the assumption of unidimensionality—namely, the test measures a single proficiency. If the UIRT procedure is applied to equate test forms that have a multidimensional data structure, the resulting equating relationships will most likely be inaccurate. When data are multidimensional, a multidimensional IRT approach may yield more accurate equating relationships.

This paper proposes an IRT observed-score equating procedure under the simple-structure multidimensional IRT (SS-MIRT) framework that can be used to equate tests that have a pre-specified set of item clusters according to a table of specifications, item types, or results from an exploratory dimensionality assessment with an oblique rotation. The specified dimensions or proficiencies are allowed to be correlated. In this paper, the SS-MIRT equating procedure is applied to a mixed-format test, in which the MC and FR items are associated with separate, yet correlated, proficiencies.

The results of a real data analysis provide the following conclusions. First, the UIRT and SS-MIRT observed score equating procedures often lead to different equating results, and the differences sometimes are substantial. These large differences might suggest that the data are not strictly unidimensional. Moreover, the results for the SS-MIRT procedure tend to be closer to the results for the equipercentile method than to the results for the UIRT procedure. Although this might be indicative of the better performance of the SS-MIRT procedure compared to the UIRT procedure, more research would be needed to draw a stronger conclusion. Especially, generalization of the results should be done with caution due to the fact that the data used in the analysis were collected under the common-item nonequivalent groups design rather than the random groups design.

The results of the simulation study show that the overall equating error for the SS-MIRT procedure is much smaller than the error for the UIRT procedure when data are multidimensional measured in terms of proficiency correlation. The SS-MIRT procedure provides adequate equating even when the proficiency correlation is very low (e.g., .5). In fact, the bias in the SS-MIRT results tends to decrease as the correlation decreases. When multidimensionality exists due to the item format effect, the SS-MIRT procedure would be preferred to the UIRT procedure. However, the UIRT procedure might produce adequate equating results if the proficiency correlation is relatively high—e.g., the scale-score results were acceptable when the correlation was .8 or higher. A future simulation study would consider other factors such as the sample size, different IRT models, and number of dimensions.

The SS-MIRT procedure can be used in conjunction with any equating design. If it is used in conjunction with the common-item nonequivalent groups design, item and proficiency parameter estimates must be put on the same scale prior to equating. It is important to remember that for a proper scaling linking, the common-item set should be representative to the total test in terms of not only the content and statistical specifications, but also the dimensional structure. In the case of a mixed-format test, a common-item set, ideally, should consist of both MC and FR items.

It should be noted that the assumption of simple structure inherent in the SS-MIRT procedure will not hold perfectly in real situations. However, it is not uncommon that a test manifests *approximate* simple structure if items in the test can be partitioned into two or more relatively distinct dimensions (Stout et al., 1996). It seems unlikely that the SS-MIRT procedure seriously distorts the equating relationship when data are not strictly simple structure demonstrating approximate simple structure. Important future research would be to examine the effect of different degrees of simple structure on equating using the SS-MIRT procedure. Note that the various useful features of the SS-MIRT procedure (e.g., easy calibration, enhanced interpretation, effective treatment of weights, etc.) become available at the cost of the relatively strong assumption of simple structure.

References

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009, September). *Equating recipes* (CASMA Monograph Number 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Brossman, B. G., & Lee, W. (2012). *Observed score and true score equating procedures for multidimensional item response theory.* Manuscript submitted for publication.

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

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

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

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

Hirsch, T. (1989). Multidimensional equating. *Journal of Educational Measurement, 26*, 337-349.

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

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

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

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*, 1-20.

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

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

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

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

Min, K. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations*. PhD thesis, Michigan State University.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating scale data* [computer program]. Chicago, IL: Scientific Software.

Oshima, T., Davey, T., & Lee, K. (2000). Multidimensional linking: four practical approaches. *Journal of Educational Measurement, 37,* 357-373.

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

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

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20,* 331-354.

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

Tate, R. (2002). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement,* 37, 329-346.

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

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

Yon, H. (2006). *Multidimensional Item Response Theory (MIRT) approaches to vertical scaling*. Unpublished doctoral dissertation, Michigan State University.

Table 1

Old-Form Conversion Table

Raw	Scale Score	Raw	Scale Score	Raw	Scale Score
0	0.00	42	24.78	84	45.03
1	0.00	43	25.28	85	45.59
2	0.00	44	25.74	86	46.13
3	0.00	45	26.24	87	46.69
4	0.00	46	26.77	88	47.32
5	0.00	47	27.24	89	47.88
6	0.00	48	27.72	90	48.42
7	0.00	49	28.21	91	49.01
8	0.00	50	28.68	92	49.55
9	0.00	51	29.13	93	50.10
10	0.00	52	29.59	94	50.68
11	0.00	53	30.06	95	51.25
12	0.00	54	30.52	96	51.86
13	1.12	55	30.99	97	52.52
14	3.62	56	31.47	98	53.15
15	5.60	57	31.95	99	53.78
16	7.24	58	32.43	100	54.51
17	8.49	59	32.90	101	55.27
18	9.23	60	33.37	102	56.01
19	9.92	61	33.82	103	56.69
20	10.84	62	34.28	104	57.35
21	11.81	63	34.76	105	58.00
22	12.71	64	35.25	106	58.59
23	13.60	65	35.73	107	59.26
24	14.45	66	36.19	108	60.13
25	15.20	67	36.66	109	61.08
26	15.89	68	37.15	110	62.03
27	16.55	69	37.65	111	62.89
28	17.16	70	38.13	112	63.56
29	17.74	71	38.59	113	64.68
30	18.36	72	39.08	114	66.27
31	18.94	73	39.56	115	67.75
32	19.49	74	40.06	116	68.87
33	20.08	75	40.56	117	69.37
34	20.63	76	41.03	118	70.00
35	21.18	77	41.53	119	70.00
36	21.72	78	42.05	120	70.00
37	22.23	79	42.56	121	70.00
38	22.76	80	43.06	122	70.00
39	23.25	81	43.55	123	70.00
40	23.74	82	44.05	124	70.00
41	24.26	83	44.53		

Table 2

Actual and Fitted Frequency Distributions

	Mean	S.D.	Skew	Kurt
New Form				
Actual	62.299	22.642	-0.093	2.123
UIRT	62.689	22.322	0.088	2.254
SS-MIRT	62.949	22.462	0.065	2.266
Old Form				
Actual	60.982	21.428	0.088	2.270
UIRT	61.900	21.541	0.225	2.341
SS-MIRT	61.462	21.372	0.229	2.365

Table 3

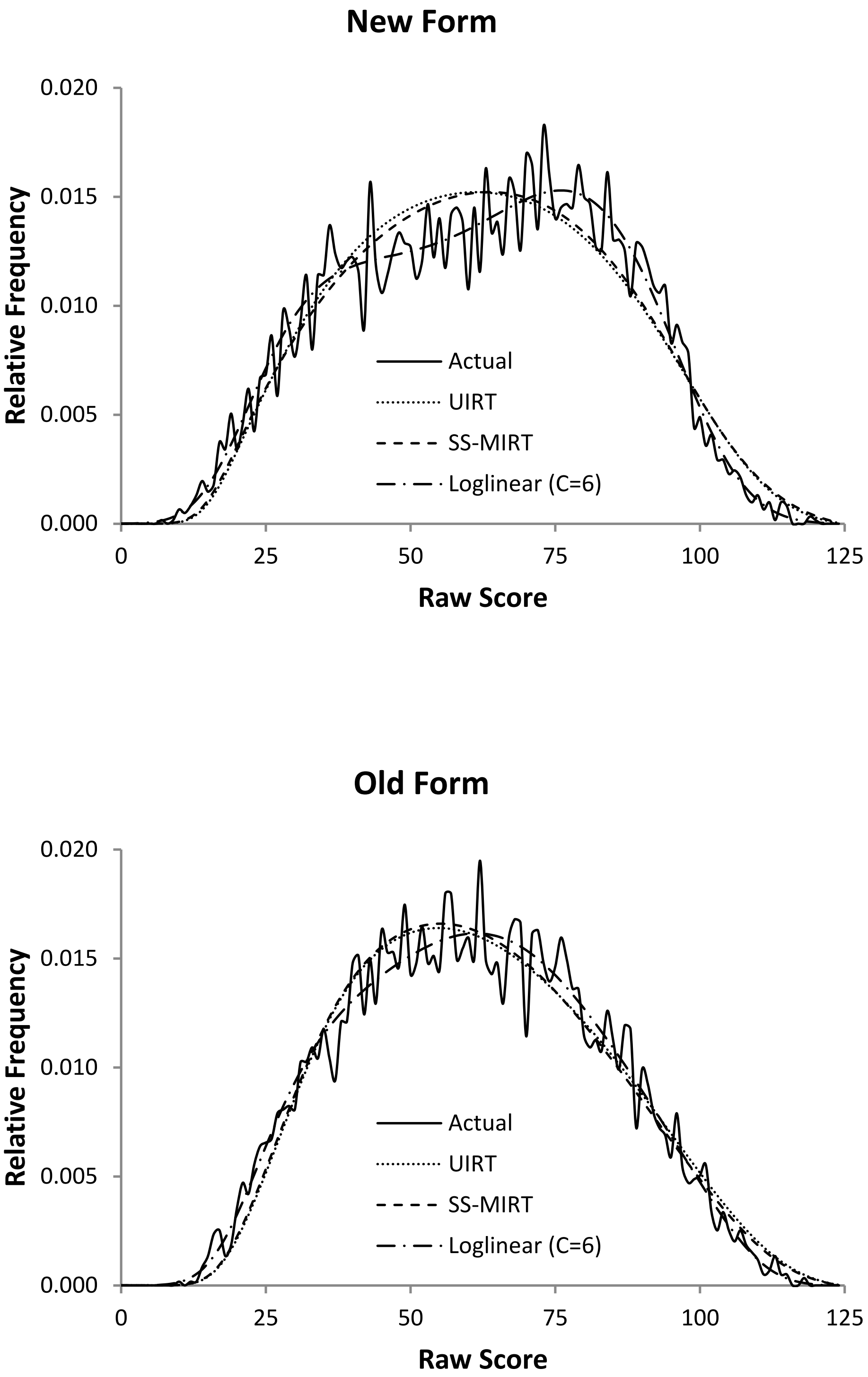
Moments for Equated Scores

	Mean	S.D.	Skew	Kurt
Raw Scores				
Old Form	60.982	21.428	0.088	2.270
	New Form Equated to Old Form Scale			
UIRT	61.532	21.725	0.027	2.171
SS-MIRT	60.852	21.391	0.050	2.179
Equipercentile	60.984	21.427	0.087	2.267
Unrounded Scale Scores				
Old Form	33.890	10.993	0.101	2.699
	New Form Equated to Old Form Scale			
UIRT	34.151	11.143	0.027	2.618
SS-MIRT	33.805	10.937	0.042	2.611
Equipercentile	33.888	10.999	0.096	2.711
Rounded Scale Scores				
Old Form	33.931	11.035	0.099	2.677
	New Form Equated to Old Form Scale			
UIRT	34.185	11.154	0.024	2.606
SS-MIRT	33.836	10.944	0.039	2.611
Equipercentile	33.857	10.983	0.093	2.710

Table 4

Summary Statistics for Simulation Results

	MSE	Squared Bias	Variance
Raw Scores			
ρ = .95			
UIRT	0.452	0.010	0.442
SS-MIRT	0.036	0.007	0.029
ρ = .8			
UIRT	0.364	0.013	0.351
SS-MIRT	0.206	0.005	0.201
ρ = .5			
UIRT	0.403	0.032	0.370
SS-MIRT	0.097	0.003	0.094
Unrounded Scale Scores			
ρ = .95			
UIRT	0.131	0.004	0.127
SS-MIRT	0.012	0.003	0.009
ρ = .8			
UIRT	0.110	0.006	0.104
SS-MIRT	0.062	0.002	0.060
ρ = .5			
UIRT	0.120	0.013	0.108
SS-MIRT	0.024	0.001	0.023
Rounded Scale Scores			
ρ = .95			
UIRT	0.283	0.073	0.209
SS-MIRT	0.149	0.071	0.078
ρ = .8			
UIRT	0.260	0.075	0.185
SS-MIRT	0.221	0.062	0.159
ρ = .5			
UIRT	0.275	0.083	0.192
SS-MIRT	0.153	0.061	0.092



New Form

Old Form

Figure 1. Actual and fitted frequency distributions.

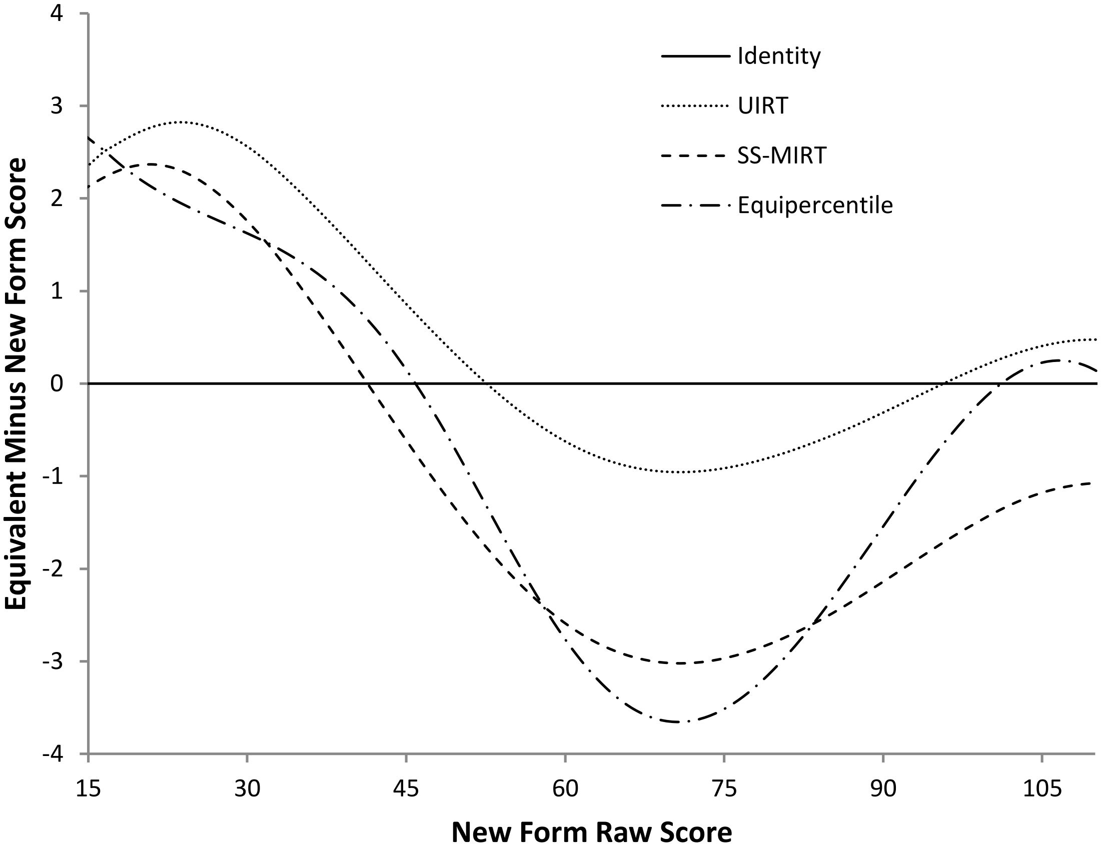


Figure 2. Raw-to-raw score equivalents.

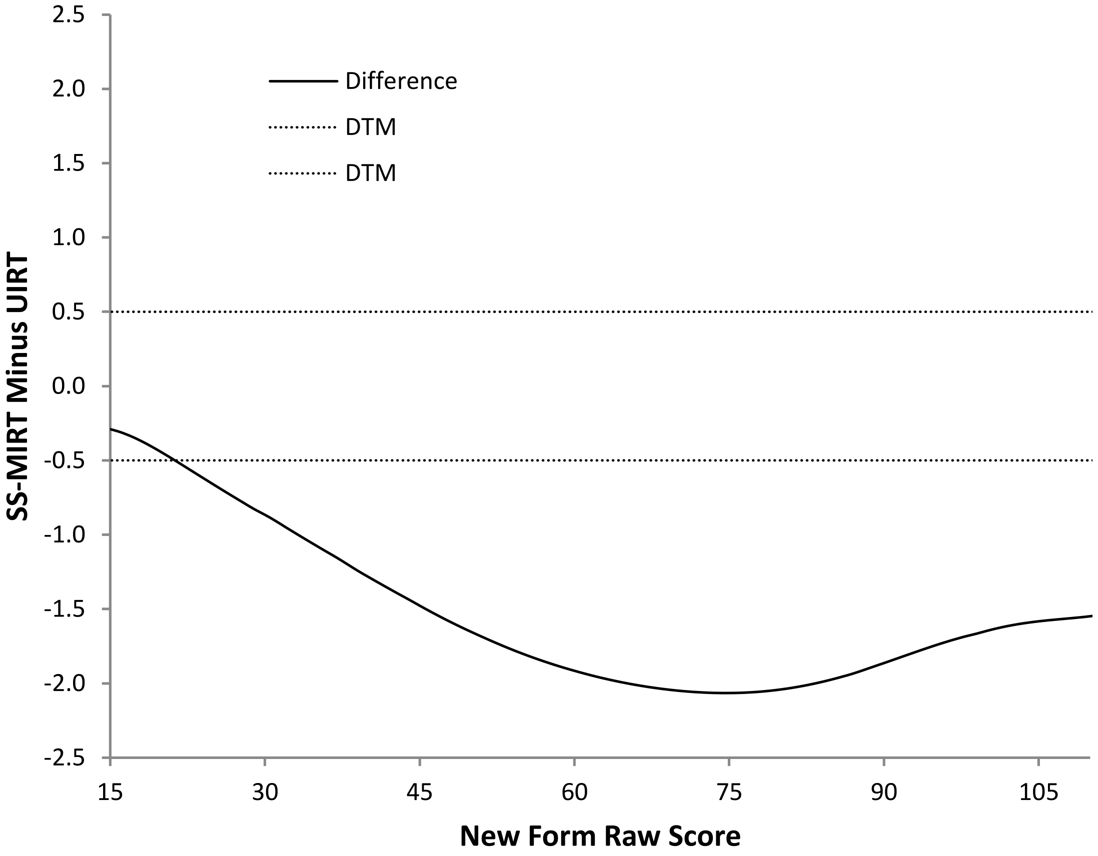


Figure 3. Differences between SS-MIRT and UIRT equated raw scores.

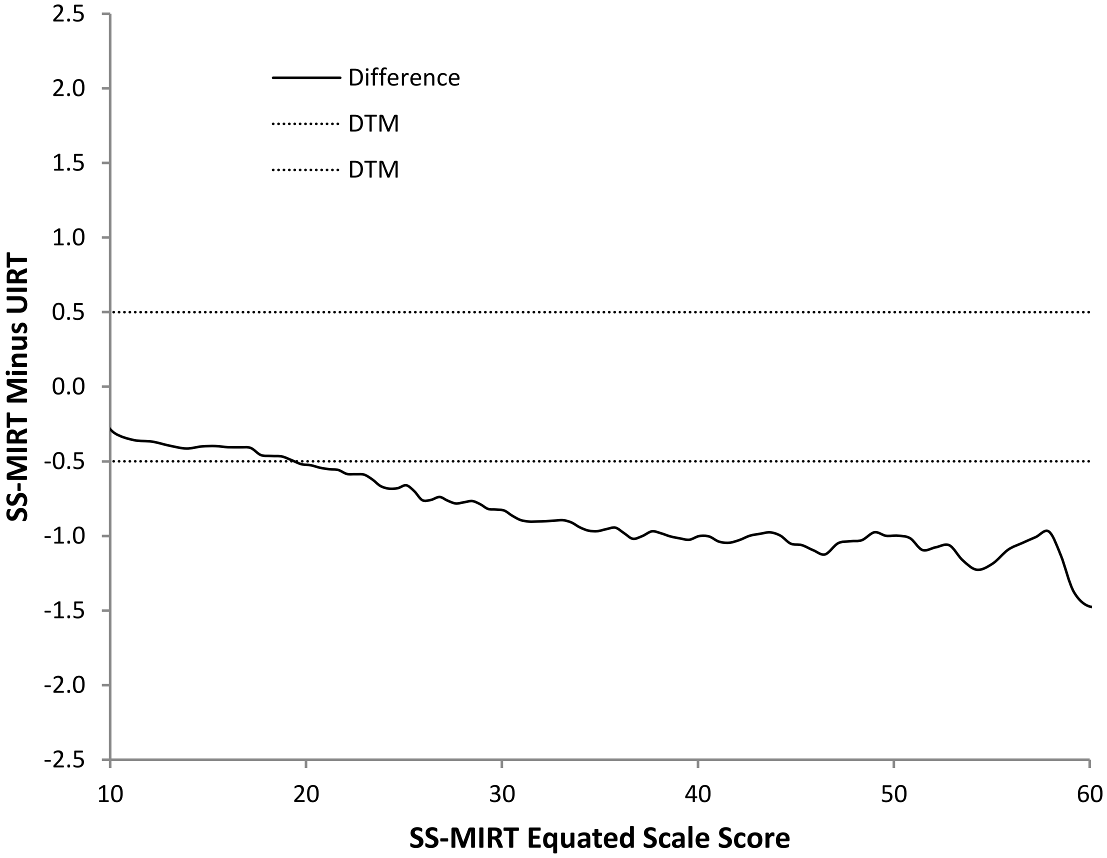


Figure 4. Differences between SS-MIRT and UIRT equated unrounded scale scores.

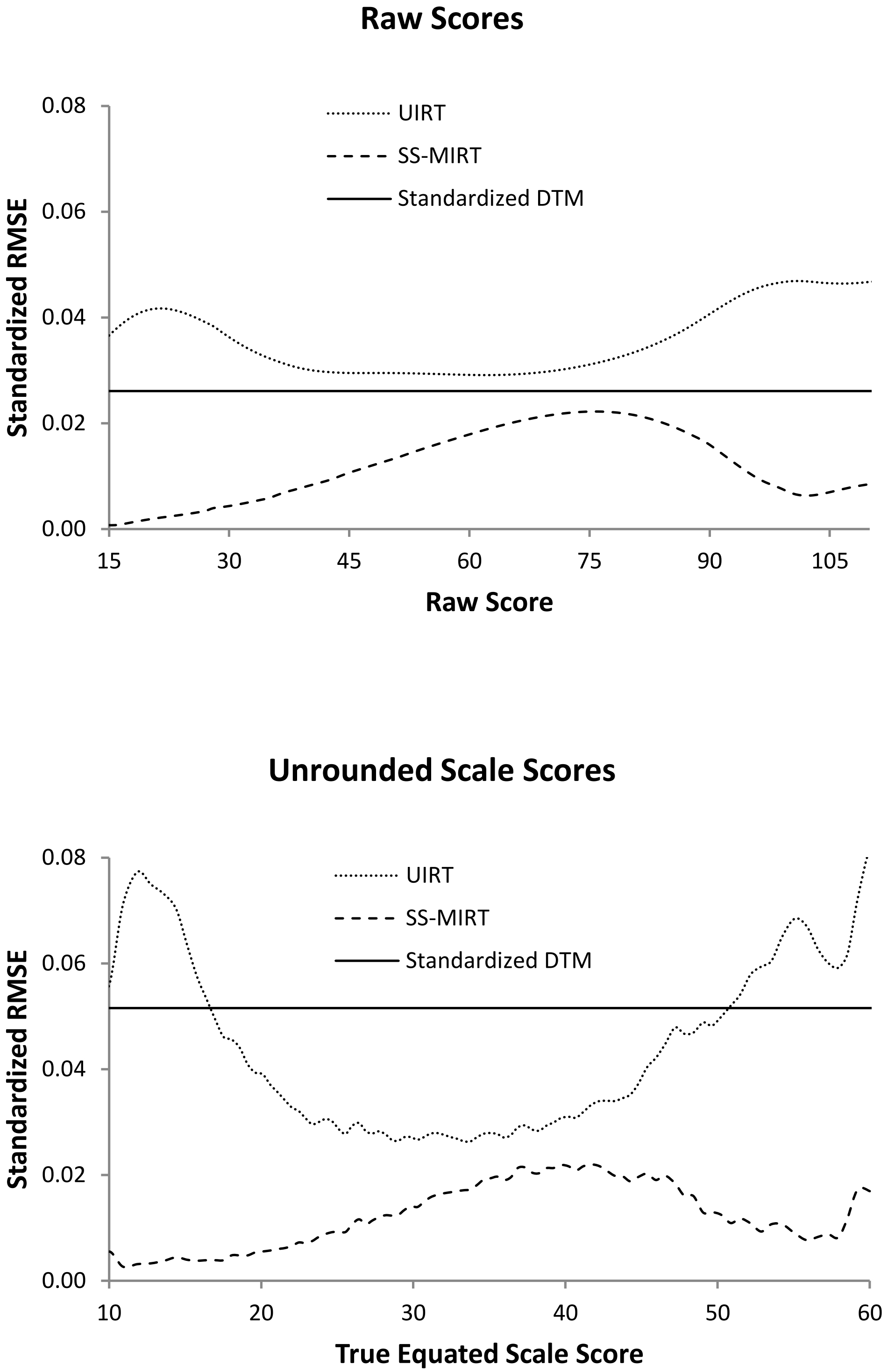


Figure 5. Conditional standardized RMSE for ρ = .5.

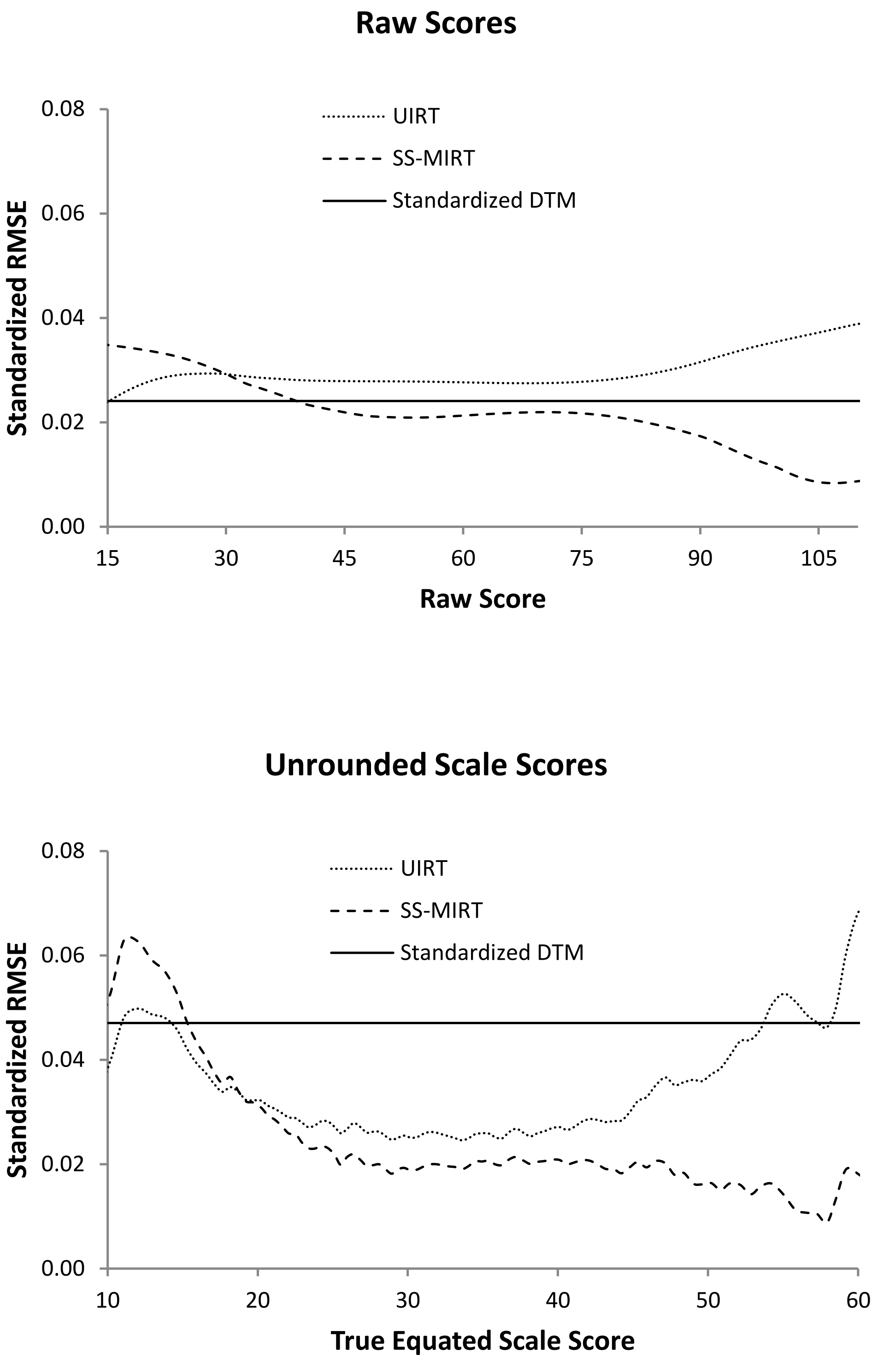
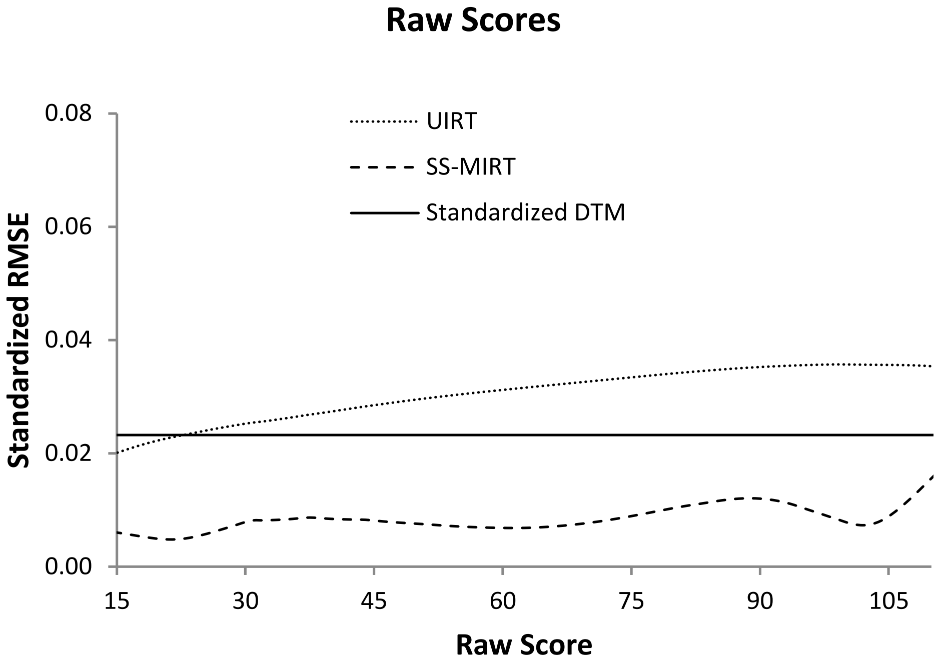


Figure 6. Conditional standardized RMSE for ρ = .8.



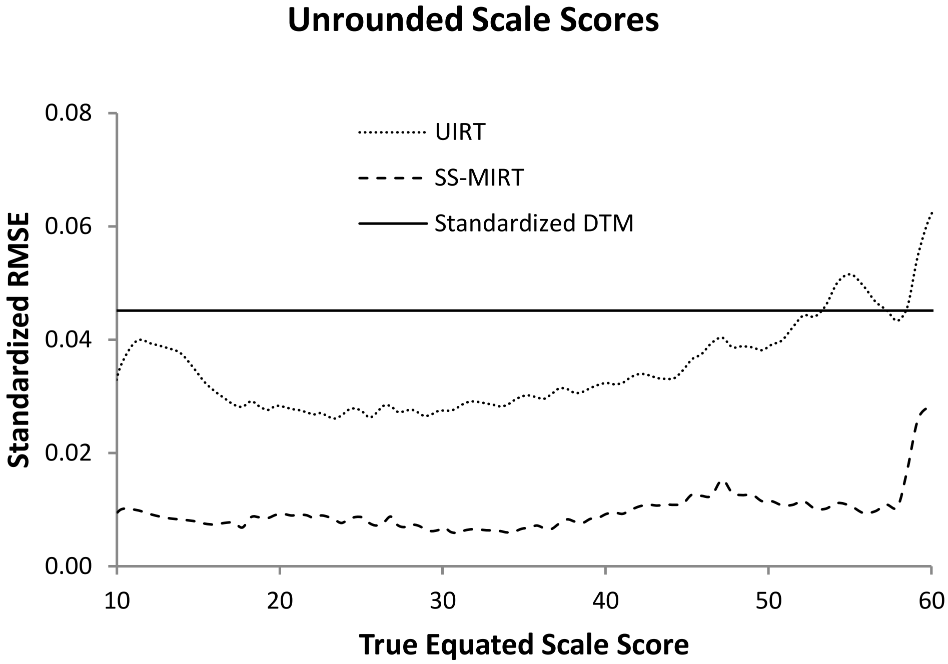


Figure 7. Conditional standardized RMSE for ρ = .95.

Chapter 6: An Investigation of IRT Item Fit Statistics for Large-Scale Mixed Format Tests

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

Abstract

This study provides an empirical investigation of various item response theory (IRT) item fit statistics when applied to large-scale mixed-format tests, Advanced Placement (AP®) Exams which consisted of multiple-choice (MC) items and different kinds of free response (FR) items. Five different item level model-data fit statistics are considered, including PARSCALE's G^2, generalized Orlando and Thissen's $S-X^2$ and $S-G^2$, and Stone's G^{2*} and χ^{2*}. Various factors that impact the performance of the statistics are investigated. As chi-squared type statistics, the five model fit statistics are sensitive to sample size. More items are flagged as misfitting by all five indices as sample size increases. Relative to PARSCALE's G^2 and Stone's statistics, the performance of the Orlando and Thissen's statistics is less affected by sample size. Meanwhile, $S-X^2$ and $S-G^2$ appear to flag fewer items as misfitting compared to the other three statistics across different test forms, sample sizes, and IRT model combinations. Only for data sets with small sample size are the numbers of misfitting items for the G^2 index located between the patterns of the Orlando's approach and the Stone's approach. In terms of consistency of the item fit indices, the statistics based on the same approach (e.g., $S-X^2$ and $S-G^2$) show higher agreement than the statistics based on different approaches (e.g., $S-G^2$ and G^{2*}). Two different IRT model mixtures used in this study appear to fit the response data well.

An Investigation of IRT Item Fit Statistics for Large-Scale Mixed Format Tests

Mixed-format tests have grown in popularity in large-scale assessments. Multiple-choice (MC) and free response (FR) items are two item types frequently used in mixed-format tests. Under item response theory (IRT), various combinations of dichotomous models (e.g., one, two, or three parameter logistical model: 1PLM, 2PLM, or 3PLM) and polytomous models (e.g., partial credit or graded response model) can be employed to analyze such mixed-format tests. When a unidimensional IRT model combination is used for a mixed-format test, violations of IRT assumptions, such as unidimensionality, might occur and lead to an inappropriate use of the chosen IRT models. To help assure the success of specific IRT applications, model-data fit and the consequences of misfit should be investigated carefully. Goodness of fit is often used as a criterion in evaluating the appropriateness of any particular model as well as detecting the violation of the use of IRT methodology.

Numerous statistical approaches have been developed to assess the fit of IRT models either for individual items or for entire test, such as Bock's χ^2 (Bock, 1972), Yen's Q_1 statistic (Yen, 1981), G^2 (McKinley & Mills, 1985), the Lagrange multiplier (LM) test (Glas & Suárez Falcón, 2003), Orlando and Thissen's (2000) $S-X^2$ and $S-G^2$ statistics, and adjusted χ^2 degrees of freedom ratios (χ^2 / dfs; Drasgow, Levine, Tsien, Williams, & Mead, 1995). This study focuses on chi-squared type item fit statistics. Specifically, five item fit statistics are considered, including PARSCALE's likelihood fit index G^2, the generalized forms of Orlando and Thissen's (2000) $S-X^2$ and $S-G^2$, and Stone's (2000) χ^{2*} and G^{2*}. This study intends to examine the performance of traditional and alternative model-data fit statistics and investigate the impact of various factors on these item fit statistics.

PARSCALE's G^2 statistic is an item fit index provided in the computer program PARSCALE (Muraki & Bock, 2003). It was selected to represent traditional chi-square type model-data fit statistics in this study. PARSCALE's G^2 statistic is a theta estimate-based index; that is, θ estimates are used in establishing the proficiency intervals in order to calculate the statistic. However, the use of the θ estimates in computing the "observed" proportions of examinees was criticized because of its dependence on the degree to which the IRT model itself is appropriate (Orlando & Thissen, 2000; Stone & Zhang, 2003). In addition, the traditional goodness-of-fit statistics were also criticized due to high sample dependence, unclearly established Type I error rates, and arbitrariness in dividing the θ continuum into discrete intervals (Orlando & Thissen, 2000; Reise, 1990). To overcome these limitations of the traditional item fit statistics, several alternative approaches have been proposed, such as the Orlando and Thissen method (2000, 2003) and the Stone method (2000). Different from the PARSCALE's G^2 statistic, Orlando and Thissen's statistics, $S-X^2$ and $S-G^2$, are based on number correct score or summed score to obtain the observed frequencies instead of the θ estimates. Observed frequencies are solely a function of the observed data, which may be superior to the model-dependent fit statistics. Orlando and Thissen's statistics were originally developed for dichotomously scored items. Later, they were generalized and applied to polytomously scored items (Kang & Chen, 2008; Schrader, Ansley, & Kim, 2004). It was argued that uncertainty in θ estimation causes the approximation of the goodness-of-fit statistics deviating from the null distribution (Stone, Mislevy, & Mazzeo, 1994). To account for the uncertainty in θ estimation, Stone (2000) suggested using posterior expectations, and developed two new fit statistics, χ^{2*} and G^{2*}.

Several studies have been performed to evaluate the performance of the traditional and alternative fit statistics, and most of them used simulated dichotomously score items. For example, Orlando and Thissen (2000) compared $S-X^2$ and $S-G^2$ with a Pearson χ^2 index (Q_1-X^2) and a likelihood ratio G^2 index (Q_1-G^2) with respect to Type I error rates and empirical power. They reported that $S-X^2$ performed consistently across different test lengths whereas $S-G^2$ performed more poorly for longer tests (80 items). The Type I error rates of $S-X^2$ were close to the nominal level; however, the other three statistics tended to reject the null hypothesis too often. In addition, $S-X^2$ was found to exhibit efficient power in the detection of item misfit. Consistent with Orlando and Thissen's findings, Stone and Zhang (2003) also concluded that the traditional statistic (Bock's χ^2 statistic) had unacceptable Type I error rates. Meanwhile, their study results demonstrated that the Stone approach exhibited adequate power in detecting misfitting items for smaller sample sizes and the Orlando and Thissen approach obtained adequate power only for a sample size of 2,000 under certain conditions.

Recently, the alternative model-data fit statistics were extended to polytomous items and mixed-format tests. Kang and Chen (2008) examined the performance of the generalized version of $S-X^2$ in assessing item-fit for the graded response model (GRM). Their simulation study results demonstrated that $S-X^2$ controlled the Type I error rates well and appropriate power was achieved with sufficient sample sizes. Chon, Lee, and Ansley (2007) examined PARSCALE's G^2 and the Orlando and Thissen's statistics using real test data. The tests contained MC items and passage-based items, and items based on the same passage were grouped together and treated as one constructed response item. Their study showed that the $S-X^2$ and $S-G^2$ indices flagged fewer items as misfitting compared to the PARSCALE index. With the same type of tests, Chon, Lee, and Ansley (in press) compared the performance of PARSCALE's G^2, the Orlando and Thissen's statistics, and the Stone's statistics. They found that the performance of the statistics varied depending on the calibrated IRT model mixtures and test characteristics. In addition, the statistics based on the same approach were reported to show higher agreement in item misfit detection relative to the statistics based on different approaches. Chon, Lee, and Dunbar (2010) investigated the same five statistics as those in the study of Chon et al. (in press) but using simulated mixed-format data. They demonstrated that the Orlando and Thissen method was the sensible and efficient choice for assessing model fit for mixed-format tests especially for short tests, and the Stone method exhibited inflated Type I error rates under certain conditions.

A review of the literature shows that very few studies have been performed to compare the traditional and alternative statistics based on real mixed-format tests. The previous literature used either simulated data or real test data in which passage-based MC items were scored as polytomous items. In practice, the mostly commonly used free-response questions are open-ended. Thus, the concern in the present study is how these statistics perform when applied to real mixed-format tests. The primary purpose of this study is to investigate the impact of various test conditions on item fit results as well as to provide empirical comparisons of traditional and alternative item fit approaches by using real mixed-format test data in real testing settings. More specifically, the purposes of this study are (a) to illustrate practical applications of various item fit statistics for data from mixed-format tests, (b) to explore the effect of multidimensionality on the item fit statistics, (c) to investigate how test characteristics such as number of examinees influence the item fit statistics, (d) to compare the performance of various model-data fit approaches in detecting misfitting items, and (e) to determine if item fit results vary according to the choice of IRT model mixtures used for calibration.

Method

Data

The data used to illustrate the selected model-data fit approaches were from Advanced Placement (AP®) examinations including Biology, Environmental Science, French Language, English Language, World History, and European History. These assessments are from the following three broader subject areas: science, language, and history. The use of different subject areas is intended to investigate whether findings are consistent across subject areas. Multiple parallel forms of some of the tests were used in this study, and item responses from a total of 13 test forms were used. The information about the 13 test forms are summarized in Table 1. All of the tests used in this study were mixed-format tests containing both MC and FR items. The English Language test was the shortest test containing 55 items (52 MC items and 3 FR items), and the other tests had at least 73 items.

The AP® exams were originally administered using formula scoring. Test takers were informed that a fraction of a point would be deducted if an MC item was answered incorrectly, and consequently, there were a substantial amount of missing responses on MC items. In the current study, number-correct scoring was used for MC items. Only those examinees who responded to at least 80% of the MC items were included in this study. In addition, for those examinees included, a two-way imputation procedure (Bernaards & Sijtsma, 2000; Sijtsma & van der Ark, 2003) was conducted to impute the responses for MC items with omitted responses. The actual sample sizes after imputation for each form are provided in Table 1.

In the current study, MC items were scored dichotomously (0/1). FR items were scored polytomously, and the item score of each FR item ranged from zero to K (number of categories minus one). For each test form used in this study, a weight of 1 was assigned to each item and used to compute total summed score for each examinee. It is important to note that, although AP® exams were used for analyses in this study the data were modified in such a way that the characteristics of the data no longer represented data from the operational AP® exam data. Consequently, generalizations of the results and findings from this study should not be made to operational AP® exams.

Assessment of Dimensionality

Unidimensionality is a critical assumption when unidimensional IRT models are used. Thus, to understand the appropriateness of the use of unidimensional IRT models, it is desirable to accurately assess the test structure prior to applying unidimensional models. Two approaches were used in this study to assess of dimensionality for each test form.

Disattenuated correlation between MC and FR sections. The disattenuated correlation between MC and FR section scores indicates the extent of construct equivalence which is closely related to dimensionality. A high disattenuated correlation tends to indicate that both item types are assessing the same dimension, whereas a low correlation suggests the existence of multidimensionality. If the unidimensionality assumption does not hold, the use of unidimensional IRT models may not be appropriate, and it is expected that item fit statistics will detect more misfitting items. When there is a strong indication of multidimensionality, separate calibration for the MC and FR items should be considered.

Disattenuated correlation between MC and FR section scores was computed using the following formula:

$$\rho_{T_M T_F} = \rho_{MF} / \sqrt{\rho_{MM'} \times \rho_{FF'}}, \tag{1}$$

where ρ_{MF} is the Pearson correlation between MC section scores (number-correct scores) and FR section scores (summed scores); and $\rho_{MM'}$ and $\rho_{FF'}$ are the coefficient α reliabilities for MC section and FR section, respectively. Equation 1 is stated in terms of parameters. For computational purposes, the parameters were replaced by estimates.

Linear factor analysis. Besides using disattenuated correlations between MC and FR section scores as an indicator of dimensionality, linear factor analysis based on polychoric correlations was also conducted to assess the dimensional structure for each test form. Three criteria for unidimensionality were considered: (1) The first eigenvalue accounts for at least 20% of the total variance (Reckase, 1979); (2) the eigenvalue of the first factor is large compared to the second and the eigenvalue of the second factor is not much larger than any of the others (Lord, 1980); and (3) an "elbow" occurs at the second eigenvalue in the scree plot.

Item Parameter Calibration

Although various IRT model combinations can be applied to mixed-format data, only practically meaningful model combinations were considered in this study. Responses for MC items were fit using the 3PLM, whereas responses of FR items were calibrated using Samejima's (1969) graded response model (GRM) or Muraki's (1992) generalized partial credit model (GPCM). Thus, two IRT model combinations, 3PLM/GRM and 3PLM/GPCM, were used to calibrate mixed-format data in this study.

The computer program PARSCALE (Muraki & Bock, 2003) was used to estimate item parameters under each of the selected IRT model combinations. For each data set, responses of items with different item types were calibrated simultaneously. The following options were used to perform the calibration. First, LOGISTIC was used, and the parameter of SCALE was set to 1.7. Second, 40 quadrature points were used instead of the default value of 30. Third, the maximum number of EM cycles was set to 300 and the maximum number of Newton-Gauss iterations (NEWTON) was set to zero. Fourth, GPRIOR and SPRIOR options were used to set prior distributions on the pseudo-guessing parameter and the slope parameter, respectively. In addition, the GPARM option was used in the calibration of Biology 2005, Environmental Science 2004 under 3PLM/GPCM, Environmental Science 2006, French Language 2004, 2006, and 2007, European History 2005 and 2007, and English Language 2007. This GPARM option, however, was not adopted in the calibration of French Language 2005 and the three World History forms (2004, 2005, and 2006) under both 3PLM/GPCM and 3PLM/GRM, because the program had convergence problem for these four forms when the GPARM option was used. Also, GPARM was not used for Environmental Science 2004 under 3PLM/GM for the same reason.

Computation of Item Fit Statistics

This study examined the performance of five item fit indices that are applicable to mixed-format data—PARSCALE's G^2, generalized versions of Orlando and Thissen's $S-X^2$ and $S-G^2$ (abbreviated hereafter as $S-X^2$ and $S-G^2$), and Stone's χ^{2*} and G^{2*}. The performance of the five indices was examined in terms of the number of misfitting items across various test conditions and IRT model combinations.

PARSCALE's G^2 was provided by the computer program PARSCALE (Muraki & Bock, 2003). Following recommendations by Chon (2009), the number of frequency score groups was set to 21 using the PARSCALE command "ITEMFIT=21" when calculating G^2. The SAS macro IRTFIT (Bjorner, Smith, Stone, & Sun, 2007) was used to compute $S-X^2$, $S-G^2$, χ^{2*}, and G^{2*}. A nominal α of .05 was used in this study. An item was flagged as misfitting if its significant level (i.e., *p*-value) for the investigated fit index was less than .05.

Factors of Investigation

Sample size. Three test forms were used to investigate the effect of sample size on the model-data fit statistics, including Environmental Science 2004, European History 2005, and French Language 2004. In order to vary sample sizes, subsets of response data with different sample sizes (*N*=1,000, 3,000, and 5,000) were randomly sampled from the original data sets for a total of nine subsets. For each test form, the five indices were calculated under four different sample size conditions (1,000, 3,000, 5,000, and the whole population) and then compared in terms of the number of misfitting items. $S-X^2$ and $S-G^2$ were calculated using a minimum expected cell frequency of 10. In addition, the nine data sets (3 test forms $\times 3$ sample sizes) were also used to study the effect of minimum expected cell frequency on the performance of $S-X^2$ and $S-G^2$ statistics, which was discussed in the later session.

IRT model mixture. Two IRT model combinations, 3PLM/GPCM and 3PLM/GRM (as discussed in "IRT Parameter Calibration"), were considered in this study. The five fit statistics were computed and compared under each IRT model combination condition. The fitting of the IRT model mixtures was evaluated not only using the nine subsets but also using the 13 original data sets.

Minimum expected cell frequency on $S-X^2$ and $S-G^2$. One of the research questions was associated with the effect of the choice of the minimum expected cell frequency on the model-data fit assessments of $S-X^2$ and $S-G^2$. Two steps were conducted to investigate this research question. In the first phase, the performance of the two fit indices on different minimum expected cell frequencies (1, 3, 5, 10, and 20) was examined using the nine subsets of response data under each selected IRT model combination. The data sets used were the same as those in the study of sample size: 3 forms (Environmental Science 2004, European History 2005, and French Language 2004) $\times 3$ sample sizes (1,000, 3,000, and 5,000). In the second phase, the item fit indices were computed using the original response data sets of the 13 forms under the two selected IRT model combinations. That is, all of the item responses of each test form were used. Different from the first phase, in the second phase, more test forms were used, and these test forms had much more examinees. The performance of $S-X^2$ and $S-G^2$ with different minimum expected cell frequencies (1, 3, 5, and 10) was studied with respect to the number of misfitting items under each IRT model combination. Note that the condition of minimum expected cell frequencies of 20 was not considered in the second phase, because the analyses in the first phase showed that the results for using 20 as the minimum expected cell frequency did not deviate much from those obtained using 10 (see Table 6).

Because real data were used in this study, the true item parameters were unknown. It was impossible to conclude which item fit index performed the best. To better understand the behavior of the five indices, findings from the present study were compared with the results provided in previous studies (e.g., Chon et al., 2010; Chon et al., in press).

Results

Assessment of Dimensionality

To understand the appropriateness of the use of unidimensional IRT models, the dimensional structure of the mixed-format data were examined to assess whether different item types (MC and FR) measure the same construct.

Dimensionality was first evaluated using disattenuated MC and FR correlation. The disattenuated correlations for the 13 test forms used in the study are presented in Table 1. Most of the forms had correlations greater than .87. However, the disattenuated MC and FR correlation of the English Language test form was only .75, and this low correlation was indicative of MC and FR measuring somewhat different constructs.

In addition, linear factor analysis was conducted to further study the dimensionality of the test forms. Table 2 lists the eigenvalues of the first five factors/dimensions and their corresponding explained total variances in percentages. The first eigenvalues accounted for at least 20% of the total variance, and they were considerably greater than the second eigenvalues for all the test forms. The second eigenvalues were not much larger than any of the others in most forms except Environmental Science 2004 and 2006, and French Language 2006. These three forms appeared to have some degree of multidimensionality because the eigenvalues for the second factors were slightly larger than any of the remaining factors. In addition, scree plots were examined as shown in Figures 1 to 3. An “elbow” occurred at the second eigenvalue for most of the test forms except Environmental Science 2004 and 2006. The scree plots indicated that the two Environmental Science forms exhibited some degree of multidimensionality.

Based on the results of the disattenuated MC and FR correlation and linear factor analysis, it appears that the English form, the Environmental Science forms (2004 and 2006), and the French Language 2006 form exhibited some degree of multidimensionality.

Effect of Sample Size on Item Fit Statistics

The effect of sample size was investigated under both 3PLM/GPCM and 3PLM/GRM with the use of three test forms from different subject areas: Environmental Science 2004, European History 2005, and French Language 2004. In addition to using the whole population, subsets with three different sample sizes (*N*=1,000, 3,000, and 5,000) were created by randomly sampling from each of the original data sets. That is, four sample sizes (*N*=1,000, 3,000, 5,000, and the original whole population) were examined for each of the three test forms.

Tables 3 and 4 provide numbers of misfitting items detected by the five item fit statistics under different sample sizes when the item responses were calibrated using the model mixture of the 3PLM and the GPCM and the model mixture of the 3PLM and the GRM, respectively. In both tables, numbers of misfitting items for different item types are reported separately. Note that the French Language test form contained 36 FR items among which the first 30 items were fill-in items with two categories (0/1) and the other 6 items were open-ended questions with at least six categories. In both of the tables, “Fill-In” is used to represent the 30 fill-in items while “OE” is used to denote the six open-ended questions.

For all three forms, as sample size increased, the five indices tended to flag more items as misfitting for the MC section, for the FR section, and for the test as a whole under either the 3PLM/GPCM or the 3PLM/GRM. In addition, compared to G^2 and Stone’s χ^{2*} and G^{2*}, Orlando and Thissen’s $S-X^2$ and $S-G^2$ statistics appeared to be less affected by sample size because the number of misfitting items did not increase as greatly as it did with the other three indices. Furthermore, below a certain sample size, Stone’s statistics tended to identify more MC items as misfitting than did Orlando and Thissen’s statistics, though this finding did not always hold for the FR items. When considering the test as a whole, more items were identified as misfitting by Stone’s statistics than by Orlando and Thissen’s statistics.

To examine whether different model-data fit statistics lead to consistent or inconsistent results in the misfit detection, the proportions of observed agreement were calculated among the five fit indices. Table 5 lists the proportions of observed agreement among the five indices when the item responses of Environmental Science 2004 with different sample sizes were calibrated using the model mixture of 3PLM and GPCM. The proportions of observed agreement among the statistics were computed for all the three test forms, and similar magnitudes were found across the test forms and various sample sizes. The results in Table 5 are representative of the results for the other forms, and thus, the results for the other two test forms are not presented here. In general, when sample size was small, the agreement among the statistics was high (over .80). When larger sample sizes were used, statistics within the same model-fit approach (e.g., $S-X^2$ and $S-G^2$) showed noticeably higher agreement compared to other pairs from different model-fit approaches, such as $S-G^2$ and χ^{2*}. The same result was obtained when the responses of the three test forms with different sample sizes were calibrated with the 3PLM/GRM model combination.

Effect of the Choice of Minimum Expected Cell Frequency on $S-X^2$ and $S-G^2$

A crucial step for calculating chi-square type fit indices is completion of the item fit table. When empty cells or sparseness in expected frequencies in the cross tabulation of item fit exist, problems can exist in the calculation of $S-X^2$ and $S-G^2$ indices. The approach of collapsing cells is often suggested to solve the sparseness problem. To use this approach, an appropriate value should be selected as the minimum expected cell frequency.

In this study, the effect of the choice of the minimum expected cell frequency on the performance of $S-X^2$ and $S-G^2$ was investigated using nine subsets as described in the Method section. Table 6 summarizes the results of $S-X^2$ and $S-G^2$ indices in terms of number of misfitting items for the three tests each under a total of 15 conditions (3 x 5 = 15) across five minimum expected cell frequencies (1, 3, 5, 10, and 20) and three sample sizes (1,000, 3,000, and 5,000). When the minimum expected cell frequency was set to 1, more items were detected as misfitting by $S-G^2$ under the three sample size conditions no matter which model mixture was used. However, this finding was not observed for the $S-X^2$ index.

The effect of the choice of minimum expected cell frequency on $S-X^2$ and $S-G^2$ was further evaluated using the original 13 response data sets each with the whole examinee population. Four minimum expected cell frequencies (1, 3, 5, and 10) were investigated. Table 7 summarizes the numbers of misfitting items detected by the two indices across various minimum expected cell frequencies and different IRT model combinations. Under each IRT model mixture, the numbers of misfitting items detected by the $S-X^2$ index were similar across various minimum expected cell frequencies. The same finding was also observed for the $S-G^2$ index. The reason for the similar proportions of misfit across different minimum expected cell frequencies is probably due to the fact that the original data sets had large sample sizes (over

10,000 examinees) and cell collapsing might not occur even using a relatively large minimum expected cell frequency (e.g., 10).

Effect of IRT Calibration Model Combinations

Two IRT model mixtures, 3PLM/GPCM and 3PLM/GRM, were used to investigate the performance of the five item fit indices. The numbers of misfit flagged by the five indices are listed in Tables 8 and 9 for 3PLM/GPCM and 3PLM/GRM, respectively, which were obtained using the 13 test forms with the whole pool of examinees. The results in Tables 8 and 9 suggest that neither 3PLM/GPCM nor 3PLM/GRM fit the data well. However, this conclusion is misleading because the five indices are chi-squared type statistics and more misfitting items with a large sample size is solely indicative of too much power in the fit statistics (a well-known problem of a chi-square statistic) rather than the degree of model fit to the data.

Instead of using the whole pool of examinees, the fit of the two model combinations was further evaluated using subsets of the original response data sets, and the subsets being used here are the same as those for the study of sample size which were described in Method section. In Table 3, columns 4 to 6 list the numbers of items flagged as misfitting by the five indices for each selected test form with different sample sizes (*N*=1,000, 3,000, and 5,000) calibrated by 3PLM/GPCM. Table 4 is for 3PLM/GRM. From Tables 3 and 4, the two IRT model mixtures showed relatively little misfit when the sample size was relatively small (i.e., 1,000), and 3PLM/GPCM fit the data better than 3PLM/GRM in most situations according to the five indices.

Comparisons among Item Fit Statistics

As shown in Tables 3 and 4, $S-X^2$ and $S-G^2$ appeared to flag fewer items as misfitting than did χ^{2*} and G^{2*} across various sample sizes and different IRT model combinations. This finding was consistent with the conclusion in the study by Chon et al. (in press) in which, for longer tests, χ^{2*} and G^{2*} indices produced higher proportions of misfitting items compared to $S-X^2$ and $S-G^2$. This finding is also consistent with the conclusion in the study by Chon et al. (2010), where this result was due to χ^{2*} and G^{2*} indices having inflated Type I error rates. When sample size was small (e.g., 1,000), fewer items were detected as misfitting by G^2 comparing to Stone’s statistics, and the proportion of misfit for the G^2 index was located between the patterns for the Stone’s and Orlando and Thissen’s statistics. However, this trend was not found for larger sample sizes (3,000 and 5,000), and it was dependent on the test forms and the IRT model combination being used. This observation is different from the previous research where Chon et al. (in press) reported that the numbers of misfitting items detected by G^2 was always located between those of the Stone's and Orlando and Thissen's statistics.

The performance of G^2, $S-X^2$, $S-G^2$, G^{2*}, and χ^{2*} was also compared using the 13 test forms with the original whole pool of examinees. The results for the five item fit indices under the 3PLM/GPCM and the 3PLM/GRM conditions are presented in Tables 8 and 9, respectively. The five fit indices showed different patterns across subject areas and different IRT model combinations. Among the five fit indices, G^2 tended to detect the most items as misfitting except for French Language 2005 under the 3PLM/GRM, French language 2006 under the 3PLM/GPCM, French Language 2007 under both the 3PLM/GPCM and the 3PLM/GRM, and European History 2005 under the 3PLM/GPCM. Compared to the $S-X^2$ and $S-G^2$ indices, Stone's G^{2*} and χ^{2*} indices classified more items as misfitting except for World History 2005. Fewer items were flagged as misfitting by $S-X^2$ and $S-G^2$ indices than by the other three statistics except for World History 2005, which is consistent with the findings obtained when smaller sample sizes were used (see Tables 3 and 4).

The proportions of observed agreement among the five fit indices were also examined to check the consistency of the statistics in misfit detection. Generally, for all the test forms investigated in this study, two pairs of the new statistics within the same model-fit approach showed noticeably higher agreement compared to other pairs from different model-fit approaches.

Effect of Multidimensionality

The effect of multidimensionality was evaluated using only Orlando and Thissen's statistics with relatively small sample sizes (1,000 and 3,000). The five fit statistics under investigation are chi-squared type statistics and they are highly influenced by sample size. When large sample size is used, they all tend to flag more items as misfitting, which makes it difficult to evaluate how tests' multidimensionality impacts the results. Therefore, relatively small sample sizes are considered for this purpose. PARSCALE's G^2 and Stone's statistics tend to inflate Type I errors, and are sensitive to sample size. From the results in this study, Orlando and

Thissen's statistics showed less sensitivity to the change in sample sizes. Therefore, only Orlando and Thissen's statistics are used to evaluate the effect of multidimensionality.

For the three test forms in Tables 3 and 4, the percentages of items detected as misfit under sample sizes of 1,000 and 3,000 were calculated and summarized in Table 10. According to the dimensionality study results, the Environmental Science 2004 form is the only one showing multidimensionality among the three forms. From Table 10, it appears that the $S-X^2$ and $S-G^2$ indices tended to flag higher proportions of items on the multidimensional test as misfitting than for the unidimensional tests.

Discussion and Conclusions

This study investigated five different model-data fit statistics using real mixed-format data, including PARSCALE's G^2, the generalized forms of Orlando and Thissen's $S-X^2$ and $S-G^2$, and Stone's χ^{2*} and G^{2*}. The performance of the item fit indices was examined under different practically meaningful IRT model combinations (3PLM/GPCM and 3PLM/GRM), and also various factors that might influence the performance of the fit indices were investigated.

The five item fit indices of investigation are chi-squared type statistics. As expected, they flagged more items as misfitting as sample size increased no matter what calibration model combination was used. As shown in Tables 3, 4, 8, and 9, the proportions of misfit for $S-X^2$ and $S-G^2$ were generally lower than the other three statistics across various sample sizes. Thus, $S-X^2$ and $S-G^2$ appear to be less sensitive to sample size compared to Stone's statistics and PARSCALE's G^2 index. Different from the previous literature (Chon et al., in press), only for small sample size (e.g., 1,000) was it observed that the proportion of misfit for the G^2 index was located between the patterns for the Stone's and Orlando and Thissen's statistics; however, this trend was not present for large sample sizes.

The consistency of the statistics in misfit detection was evaluated using the proportions of observed agreement among the five fit indices across various sample sizes. Consistent with the findings in previous studies (Chon et al., in press), pairs of statistics based on the same approach exhibited higher agreement compared to pairs of statistics from different approaches.

Under sample sizes of 1,000, 3,000, and 5,000, more items were flagged as misfitting by $S-G^2$ when the minimum expected cell frequency was set to 1 than for other expected cell frequencies regardless of the calibration methods. However, this finding was not observed for the

$S-X^2$ statistic. When sample size became larger, there was not much difference in using 1, 3, 5, or 10 as the cell frequency considering the performance of the two indices in this study.

In practice, fit of the two IRT model combinations should be evaluated using relatively small sample sizes, because chi-squared type statistics are sensitive to sample size and results obtained by using large sample size (e.g., over 10,000) are somewhat misleading. Using data sets with sample size of 1,000, both 3PLM/GPCM and 3PLM/GRM showed very little misfit, and 3PLM/GPCM tended to fit the data better than 3PLM/GRM.

The effect of multidimensionality was evaluated using Orlando and Thissen's $S-X^2$ and $S-G^2$ statistics under relatively small sample sizes (1,000 and 3,000). The fit statistics tended to detect more item as misfitting for multidimensional tests than for unidimensional tests. However, in this study, multidimensionality is confounded with subject areas. To better understand the effect of multidimensionality, more research would be necessary.

One major contribution of this study is that it evaluates item fit statistics using real mixed-format tests that contain MC items and open-ended FR items. MC items and open-ended FR items are the used widely in mixed-format tests. Previous literature focused on using simulated data or passage-based items (not open-ended questions), and their conclusions might not be generalizable to tests containing open-ended FR items. The current study can help fill this gap.

One limitation of the study is that the AP data used were after imputation for omitted item responses. It is still not clear how imputation procedure impacts the performance of the item fit statistics. Further studies are needed to investigate the effect of imputation or with tests for which there are few omitted responses.

Item fit indices should be able to adequately identify items that misfit in a way that has practical consequences for the use of the IRT models. As chi-squared type statistics, the five item fit statistics used in this study are substantially influenced by sample size. To avoid over- or under-identifying misfitting items, criteria that are not so closely related to sample size are highly desired. In the current literature, no such criteria are available. Thus, the development of such criteria is an important area for further work in item fit statistics.

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35,* 321-364. doi: 10.1207/S15327906MBR3503_03

Bjorner, J. B., Smith, K. J., Stone, C. A., & Sun, X. (2007). *IRTFIT: A macro for item fit and local dependence tests under IRT models*. Lincoln, RI: Quality Metric, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51. doi: 10.1007/BF02291411

Chon, K. H. (2009). *An investigation of item fit statistics for mixed IRT models*. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47,* 318-338.

Chon, K. H., Lee, W., & Ansley, T. (2007, November). *Assessing IRT model-data fit for mixed format tests*. Retrieved from http://www.education.uiowa.edu/centers/casma/research-reports.aspx

Chon, K. H., Lee, W., & Ansley, T. N. (in press). An empirical investigation of methods for assessing item fit for mixed format tests. *Applied Measurement in Education.*

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting the polytomous item response theory models to multiple-choice items. *Applied Psychological Measurement*, *19*, 143-165.

Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Kang, T. & Chen, T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*, 391-406.

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

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49-57.

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

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating scale data* [computer program]. Chicago, IL: Scientific Software.

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

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

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127-137.

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

von Schrader, S., Ansley, T., & Kim, S. (2004, April). *Examination of item ift indices for polytomous item response models*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*: 505-528.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Stone, C. A., Mislevy, R. J., & Mazzeo, J. (1994, April). *Classification error and goodness-of-fit in IRT models*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Stone, C. A., & Zhang, B. (2003). Assessing goodness-of-fit of IRT Models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Table 1

Summary of Test Information

Subject Area	Test	Form	Disattenuated Correlation between MC and FR	Composite Reliability	Test Length		# of Points for Per FR Items	# of Examinees
					# of MC	# of FR		
Science	Biology	2005	0.941	0.942	98	4	10	16,185
	Environmental Science	2004	0.942	0.935	99	4	10	16,999
		2006	0.933	0.939	100	4	10	17,508
Language	French Language	2004	0.884	0.963	80	36[a]	1,9,5	12,349
		2005	0.877	0.959	79	36[a]	1,9,5	13,571
		2006	0.873	0.963	84	36[a]	1,9,5	13,403
		2007	0.881	0.959	85	36[a]	1,9,5	13,982
	English Language	2007	0.750	0.895	52	3	9	16,882
History	World History	2004	0.908	0.909	70	3	9	17,043
		2005	0.872	0.926	70	3	9	16,965
		2007	0.899	0.924	70	3	9	16,713
	European History	2005	0.883	0.920	80	3	9	16,084
		2007	0.888	0.916	80	3	9	16,799

[a]: Of the 36 FR items, the first 30 items are fills-in items. Each fills-in item has two categories. The number of points for each fills-in item is 1. The 31st FR item has a maximum score point of 9 (10 categories). Each of the remaining five FR items has a maximum score point of 5 (6 categories).

Table 2

Results of Linear Factor Analysis

		Factor Analysis					
Subject Area	Test Forms		Dimension/Factor				
			1	2	3	4	5
Science	Biology 2005	Eigenvalue	27.62	2.74	1.81	1.66	1.40
		% of Variance	27.08	2.68	1.77	1.63	1.37
	Environmental Science 2004	Eigenvalue	24.53	4.25	1.96	1.59	1.31
		% of Variance	23.81	4.13	1.90	1.55	1.27
	Environmental Science 2006	Eigenvalue	26.59	3.44	1.79	1.65	1.26
		% of Variance	25.56	3.31	1.72	1.59	1.21
Language	French Language 2004	Eigenvalue	38.49	3.33	2.42	1.94	1.53
		% of Variance	33.18	2.87	2.09	1.67	1.33
	French Language 2005	Eigenvalue	36.81	3.63	2.82	2.00	1.37
		% of Variance	32.01	3.16	2.45	1.74	1.19
	French Language 2006	Eigenvalue	37.77	4.07	2.76	2.04	1.49
		% of Variance	31.48	3.39	2.30	1.70	1.24
	French Language 2007	Eigenvalue	35.23	3.67	2.80	2.20	1.47
		% of Variance	29.12	3.04	2.32	1.82	1.22
	English Language 2007	Eigenvalue	14.37	1.60	1.48	1.20	1.08
		% of Variance	26.13	2.91	2.69	2.19	1.97
History	World History 2004	Eigenvalue	17.03	2.07	1.27	1.22	1.13
		% of Variance	23.32	2.83	1.74	1.67	1.55
	World History 2005	Eigenvalue	19.60	1.85	1.42	1.14	1.08
		% of Variance	26.85	2.53	1.95	1.56	1.47
	World History 2006	Eigenvalue	20.47	1.65	1.22	1.14	1.11
		% of Variance	28.05	2.27	1.67	1.56	1.52
	European History 2005	Eigenvalue	20.57	2.01	1.59	1.34	1.11
		% of Variance	24.78	2.42	1.92	1.61	1.34
	European History 2006	Eigenvalue	20.71	2.05	1.58	1.27	1.21
		% of Variance	24.95	2.47	1.91	1.53	1.46

Table 3

Number of Misfitting Items for Environmental Science 2004, European History 2005, and French Language 2004 under Different Sample Sizes (N=1,000, 3,000, 5,000, and the Whole Population) with 3PLM/GPCM

Test/Form	Statistics	Item Type	Number of Misfitting Items			
			Number of Examinees			
			1000	3000	5000	Whole Pop. (above 12,000)
Env. Sci 2004 (MC: 99; FR: 4; Total: 103 items)	G^2	MC	4	15	21	71
		FR	0	2	4	4
	$S\text{-}G^2$	MC	2	7	7	25
		FR	1	0	1	4
	$S\text{-}X^2$	MC	2	7	7	24
		FR	1	0	1	4
	G^{2*}	MC	9	27	31	60
		FR	0	2	4	4
	χ^{2*}	MC	16	26	32	60
		FR	0	2	4	4
European History 2005 (MC: 80; FR: 3; Total: 83 items)	G^2	MC	6	12	20	43
		FR	0	0	1	3
	$S\text{-}G^2$	MC	1	6	14	28
		FR	0	0	0	3
	$S\text{-}X^2$	MC	1	6	13	28
		FR	0	0	0	3
	G^{2*}	MC	13	15	27	46
		FR	0	1	1	3
	χ^{2*}	MC	14	17	27	45
		FR	1	1	1	3
French Language 2004 (MC: 80; Fill-In: 30; OE: 6; Total: 116 items)	G^2	MC	4	37	43	65
		Fills-In	2	17	22	28
		OE	1	6	6	6
	$S\text{-}G^2$	MC	2	2	6	10
		Fills-In	0	1	3	5
		OE	0	0	0	2
	$S\text{-}X^2$	MC	1	2	6	11
		Fills-In	0	1	2	5
		OE	0	0	0	2
	G^{2*}	MC	10	14	23	40
		Fills-In	4	6	12	20
		OE	0	2	3	6
	χ^{2*}	MC	15	15	22	43
		Fills-In	3	6	11	18
		OE	1	4	5	6

Table 4

Number of Misfitting Items for Environmental Science 2004, European History 2005, and French Language 2004 under Different Sample Sizes (N=1,000, 3,000, 5,000, and the Whole Population) with the Use of 3PLM/GRM

Test/Form	Statistics	Item Type	Number of Misfitting Items			
			Number of Examinees			
			1,000	3,000	5,000	Whole Pop. (above 12,000)
Env. Sci 2004 (MC: 99; FR: 4; Total: 103 items)	G^2	MC	9	55	92	99
		FR	1	4	4	4
	$S\text{-}G^2$	MC	2	10	14	45
		FR	1	0	0	4
	$S\text{-}X^2$	MC	2	10	14	47
		FR	1	0	0	4
	G^{2*}	MC	10	33	31	69
		FR	0	2	4	4
	χ^{2*}	MC	14	32	31	68
		FR	0	2	4	4
European History 2005 (MC: 80; FR: 3; Total: 83 items)	G^2	MC	5	14	22	53
		FR	0	1	3	3
	$S\text{-}G^2$	MC	1	6	14	31
		FR	0	0	0	3
	$S\text{-}X^2$	MC	1	6	15	30
		FR	0	0	0	3
	G^{2*}	MC	11	18	27	45
		FR	0	3	3	3
	χ^{2*}	MC	13	18	28	45
		FR	0	2	3	3
French Language 2004 (MC: 80; Fill-In: 30; OE: 6; Total: 116 items)	G^2	MC	3	32	45	70
		Fills-In	1	15	23	29
		OE	1	6	6	6
	$S\text{-}G^2$	MC	2	2	6	11
		Fills-In	0	1	3	5
		OE	0	0	0	1
	$S\text{-}X^2$	MC	1	2	6	11
		Fills-In	0	1	2	5
		OE	0	0	0	1
	G^{2*}	MC	10	16	22	41
		Fills-In	4	7	12	19
		OE	0	1	4	6
	χ^{2*}	MC	15	18	21	43
		Fills-In	3	7	11	17
		OE	1	2	2	6

Table 5

Proportions of Observed Agreement among the Five Item Fit Statistics for Environmental Science 2004 under Different Sample Sizes and 3PLM/GPCM

Sample Size		G^2	$S\text{-}G^2$	$S\text{-}X^2$	G^{2*}	χ^{2*}
1,000	G^2	1.00	.97	.97	.91	.86
	$S\text{-}G^2$	.97	1.00	1.00	.90	.83
	$S\text{-}X^2$	.97	1.00	1.00	.90	.83
	G^{2*}	.91	.90	.90	1.00	.89
	χ^{2*}	.86	.83	.83	.89	1.00
3,000	G^2	1.00	.90	.90	.84	.86
	$S\text{-}G^2$	.90	1.00	1.00	.79	.80
	$S\text{-}X^2$	.90	1.00	1.00	.79	.80
	G^{2*}	.84	.79	.79	1.00	.99
	χ^{2*}	.86	.80	.80	.99	1.00
5,000	G^2	1.00	.83	.83	.88	.87
	$S\text{-}G^2$	.83	1.00	1.00	.74	.73
	$S\text{-}X^2$	.83	1.00	1.00	.74	.73
	G^{2*}	.88	.74	.74	1.00	.97
	χ^{2*}	.87	.73	.73	.97	1.00
16,999	G^2	1.00	.55	.53	.80	.80
	$S\text{-}G^2$	.55	1.00	.99	.66	.66
	$S\text{-}X^2$	.53	.99	1.00	.65	.65
	G^{2*}	.80	.66	.65	1.00	1.00
	χ^{2*}	.80	.66	.65	1.00	1.00

Table 6

Effect of Minimum Expected Cell Frequency on S-G^2 and S-X^2 Using Environmental Science 2004, French Language 2004, and European History 2005 with Sample Sizes of 1,000, 3,000, and 5,000 under both 3PLM/GPCM and 3PLM/GRM

Sample Size	Tests/Forms	Model	Number of Misfitting Items									
			S-G^2					S-X^2				
			Minimum Expected Cell Frequency					Minimum Expected Cell Frequency				
			1	3	5	10	20	1	3	5	10	20
1,000	Env. Sci. 2004 (103 items)	3PLM/GPCM	13	4	3	3	3	5	2	2	3	2
		3PLM/GRM	13	4	2	3	1	4	3	2	3	1
	French 2004 (116 items)	3PLM/GPCM	7	2	3	2	1	3	1	1	1	1
		3PLM/GRM	7	2	3	2	1	3	1	1	1	1
	Eur. His. 2005 (83 items)	3PLM/GPCM	3	1	0	1	2	1	0	0	1	2
		3PLM/GRM	3	1	0	1	2	1	0	0	1	2
3,000	Env. Sci. 2004 (103 items)	3PLM/GPCM	12	9	10	7	8	9	9	6	7	8
		3PLM/GRM	14	14	12	10	12	12	11	10	10	12
	French 2004 (116 items)	3PLM/GPCM	10	7	4	3	4	7	6	3	3	3
		3PLM/GRM	11	7	4	3	4	7	6	3	3	3
	Eur. His. 2005 (83 items)	3PLM/GPCM	12	9	7	6	6	8	6	6	6	6
		3PLM/GRM	12	8	7	6	5	8	6	6	6	5
5,000	Env. Sci. 2004 (103 items)	3PLM/GPCM	15	10	8	8	10	10	8	8	8	10
		3PLM/GRM	24	17	16	14	17	18	14	14	14	17
	French 2004 (116 items)	3PLM/GPCM	13	11	10	9	8	10	10	10	8	6
		3PLM/GRM	13	11	10	9	8	10	10	10	8	6
	Eur. His. 2005 (83 items)	3PLM/GPCM	17	15	14	14	13	13	12	13	13	13
		3PLM/GRM	18	15	14	14	14	13	13	15	15	14

Table 7

Effect of Minimum Expected Cell Frequency on S-G^2 and S-X^2 Using the Whole Population

Tests/Forms	Model	Number of Misfitting Items							
		S-G^2				*S-X^2*			
		Minimum Expected Cell Frequency				Minimum Expected Cell Frequency			
		1	3	5	10	1	3	5	10
Biology 2005 (102 items)	3PLM/GPCM	54	51	52	49	54	51	51	49
	3PLM/GRM	56	53	50	51	56	52	52	52
Environmental Science 2004 (103 items)	3PLM/GPCM	35	34	33	29	29	32	32	28
	3PLM/GRM	52	47	47	49	51	49	50	51
Environmental Science 2006 (104 items)	3PLM/GPCM	45	37	39	37	41	35	34	33
	3PLM/GRM	48	38	41	37	46	38	39	37
French 2004 (116 items)	3PLM/GPCM	18	15	15	17	20	19	18	18
	3PLM/GRM	18	16	16	17	18	16	16	17
French 2005 (115 items)	3PLM/GPCM	35	31	30	28	36	36	32	29
	3PLM/GRM	36	38	36	34	39	38	36	36
French 2006 (120 items)	3PLM/GPCM	40	37	35	34	39	35	34	33
	3PLM/GRM	43	39	37	37	40	35	36	36
French 2007 (121 items)	3PLM/GPCM	56	52	50	43	52	52	48	43
	3PLM/GRM	57	52	52	48	52	51	50	44
English 2007 (55 items)	3PLM/GPCM	26	25	25	26	27	27	28	26
	3PLM/GRM	31	31	32	33	32	30	32	33
World History 2004 (73 items)	3PLM/GPCM	33	31	28	29	32	32	28	31
	3PLM/GRM	38	39	37	35	40	42	38	37
World History 2005 (73 items)	3PLM/GPCM	48	47	47	47	49	50	49	50
	3PLM/GRM	53	53	52	53	54	55	55	55
World History 2006 (73 items)	3PLM/GPCM	26	26	26	24	25	24	24	24
	3PLM/GRM	32	30	31	31	30	30	31	31
European History 2005 (83 items)	3PLM/GPCM	33	32	33	31	30	31	33	31
	3PLM/GRM	39	37	37	34	39	37	39	33
European History 2007 (83 items)	3PLM/GPCM	39	36	34	33	33	35	33	32
	3PLM/GRM	40	37	35	34	36	38	36	33

Table 8

Number of Misfitting Items for the 13 Data Sets Based on Five Item Fit Statistics under the 3PLM/GPCM Using the Whole Population

Test/Form	Model	Item Type	Number of Misfitting Items				
			G^2[a]	$S\text{-}G^2$[b]	$S\text{-}X^2$[b]	G^{2*}	χ^{2*}
Biology 2005 (MC:98; FR:4)	3PLM/GPCM	MC	88	45	45	65	65
		FR	4	4	4	4	4
Environmental Science 2004 (MC:99; FR:4)	3PLM/GPCM	MC	71	25	24	60	60
		FR	4	4	4	4	4
Environmental Science 2006 (MC:100; FR:4)	3PLM/GPCM	MC	100	33	29	68	67
		FR	4	4	4	4	4
French Language 2004 (MC:80; Fill-In:30; OE:6)	3PLM/GPCM	MC	65	10	11	40	43
		Fills-In	28	5	5	20	18
		OE	6	2	2	6	6
French Language 2005 (MC:79 Fill-In:30; OE:6)	3PLM/GPCM	MC	26	17	18	47	47
		Fills-In	16	11	11	18	19
		OE	6	0	0	6	6
French Language 2006 (MC:84; Fill-In:30; OE:6)	3PLM/GPCM	MC	33	17	18	43	43
		Fills-In	24	13	11	26	26
		OE	6	4	4	6	6
French Language 2007 (MC:85; Fill-In:30; OE:6)	3PLM/GPCM	MC	33	23	23	45	44
		Fills-In	22	19	18	24	24
		OE	6	1	2	6	6
English Language 2007 (MC:52; FR:3)	3PLM/GPCM	MC	50	23	23	35	35
		FR	3	3	3	3	3
World History 2004 (MC:70; FR:3)	3PLM/GPCM	MC	68	26	28	39	40
		FR	3	3	3	3	3
World History 2005 (MC:70; FR:3)	3PLM/GPCM	MC	70	44	47	45	45
		FR	3	3	3	3	3
World History 2006 (MC:70; FR:3)	3PLM/GPCM	MC	63	22	22	47	48
		FR	3	2	2	3	3
European History 2005 (MC:80; FR:3)	3PLM/GPCM	MC	43	28	28	46	45
		FR	3	3	3	3	3
European History 2007 (MC:80; FR:3)	3PLM/GPCM	MC	48	31	30	44	44
		FR	3	2	2	2	2

[a]: The number of frequency score groups was set to 21 for the calculation of G^2.

[b]: The minimum expected cell frequency was set to 10 for the calculation of $S\text{-}G^2$ and $S\text{-}X^2$ indices.

Table 9

Number of Misfitting Items for the 13 Data Sets Based on Five Item Fit Statistics under the 3PLM/GRM Using the Whole Population

Test/Form	Model	Item Type	Number of Misfitting Items				
			$G^{2\,a}$	$S\text{-}G^{2\,b}$	$S\text{-}X^{2\,b}$	G^{2*}	χ^{2*}
Biology A (MC:98; FR:4)	3PLM/ GRM	MC	95	47	48	67	67
		FR	4	4	4	4	4
Environmental Science 2004 (MC:99; FR:4)	3PLM/GPCM	MC	99	45	47	69	68
		FR	4	4	4	4	4
Environmental Science 2006 (MC:100; FR:4)	3PLM/GPCM	MC	94	34	34	67	67
		FR	4	3	3	4	4
French Language 2004 (MC:80; Fill-In:30; OE:6)	3PLM/GPCM	MC	70	11	11	41	43
		Fills-In	29	5	5	19	17
		OE	6	1	1	6	6
French Language 2005 (MC:79 Fill-In:30; OE:6)	3PLM/GPCM	MC	35	18	20	49	45
		Fills-In	19	11	11	18	19
		OE	6	5	5	6	6
French Language 2006 (MC:84; Fill-In:30; OE:6)	3PLM/GPCM	MC	49	19	20	43	43
		Fills-In	26	13	11	26	25
		OE	6	5	5	6	6
French Language 2007 (MC:85; Fill-In:30; OE:6)	3PLM/GPCM	MC	40	24	23	45	46
		Fills-In	24	19	17	24	24
		OE	6	5	4	6	6
English Language 2007 (MC:52; FR:3)	3PLM/GPCM	MC	50	30	30	34	34
		FR	3	3	3	3	3
World History 2004 (MC:70; FR:3)	3PLM/GPCM	MC	70	32	34	40	40
		FR	3	3	3	3	3
World History 2005 (MC:70; FR:3)	3PLM/GPCM	MC	70	50	52	46	46
		FR	3	3	3	3	3
World History 2006 (MC:70; FR:3)	3PLM/GPCM	MC	70	28	28	48	48
		FR	3	3	3	3	3
European History 2005 (MC:80; FR:3)	3PLM/GPCM	MC	53	31	30	45	45
		FR	3	3	3	3	3
European History 2007 (MC:80; FR:3)	3PLM/GPCM	MC	75	31	30	44	44
		FR	3	3	3	3	3

Table 10

100×Proportion of Misfitting Items Detected by Orlando and Thissen's Statistics Under Sample Sizes of 1,000 and 3,000

Test Form	Model	Sample Size = 1,000		Sample Size = 3,000	
		S-G^2	*S-X*^{2}	*S-G*^{2}	*S-X*^{2}
Env. Sci 2004 (Total: 103 items)	*3PLM/GPCM*	2.913	2.913	6.796	6.796
European History 2005 (Total: 83 items)	*3PLM/GPCM*	1.205	1.205	7.229	7.229
French Language 2004 (Total: 116 items)	*3PLM/GPCM*	1.724	.862	.026	.026
Env. Sci 2004 (Total: 103 items)	*3PLM/GRM*	2.913	2.913	9.709	9.709
European History 2005 (Total: 83 items)	*3PLM/GRM*	1.205	1.205	7.229	7.229
French Language 2004 (Total: 116 items)	*3PLM/GRM*	1.724	.862	.026	.026

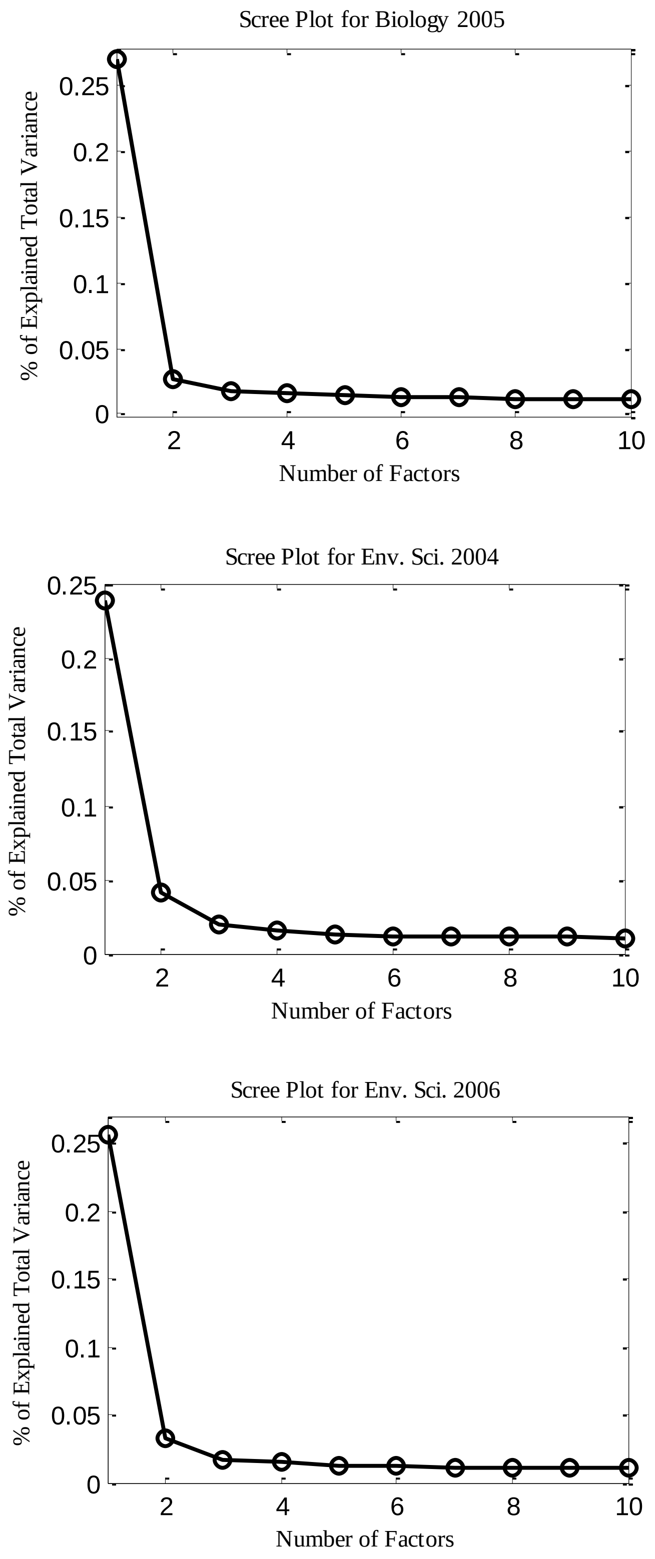


Figure 1. Scree plots for science test forms.

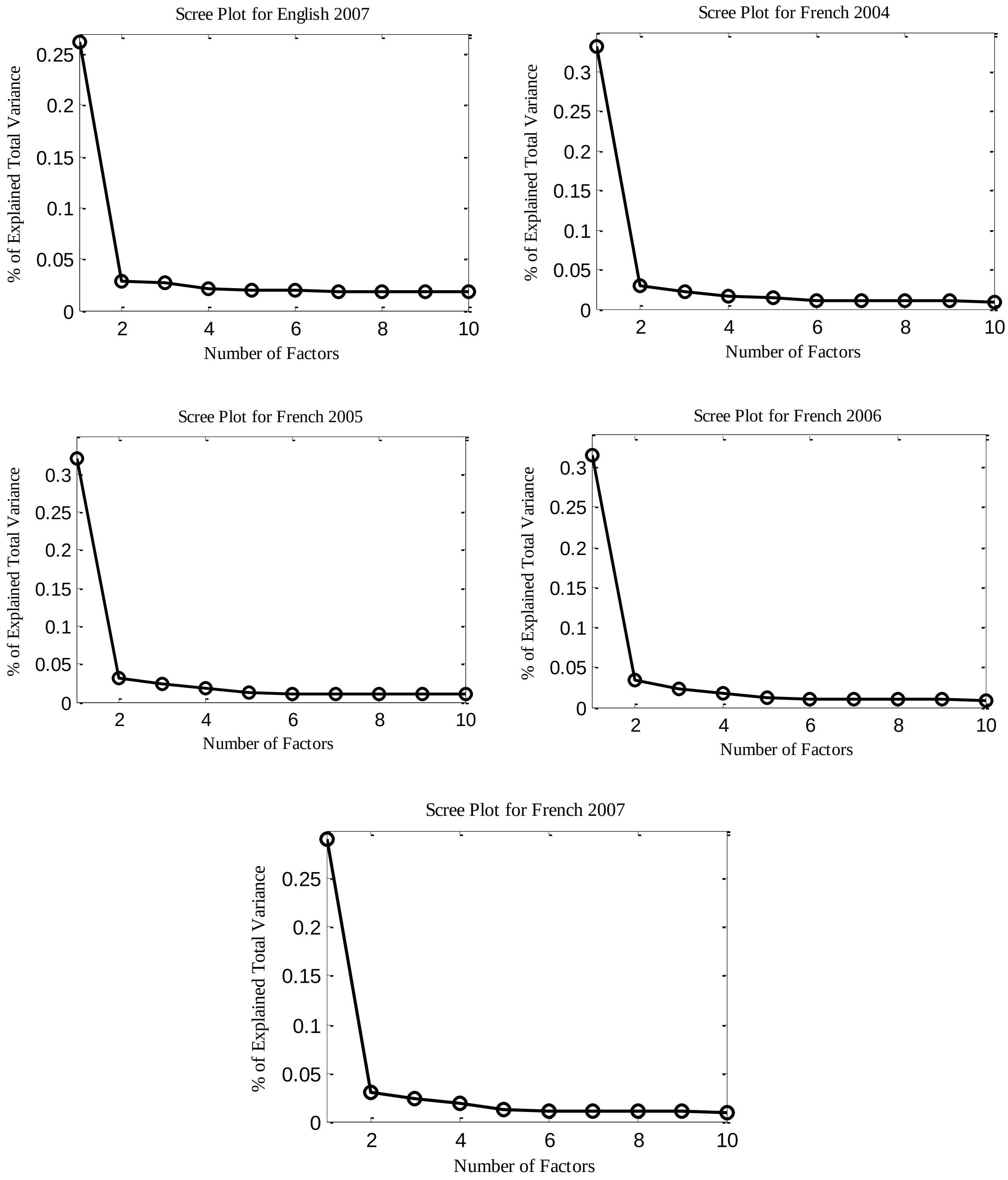
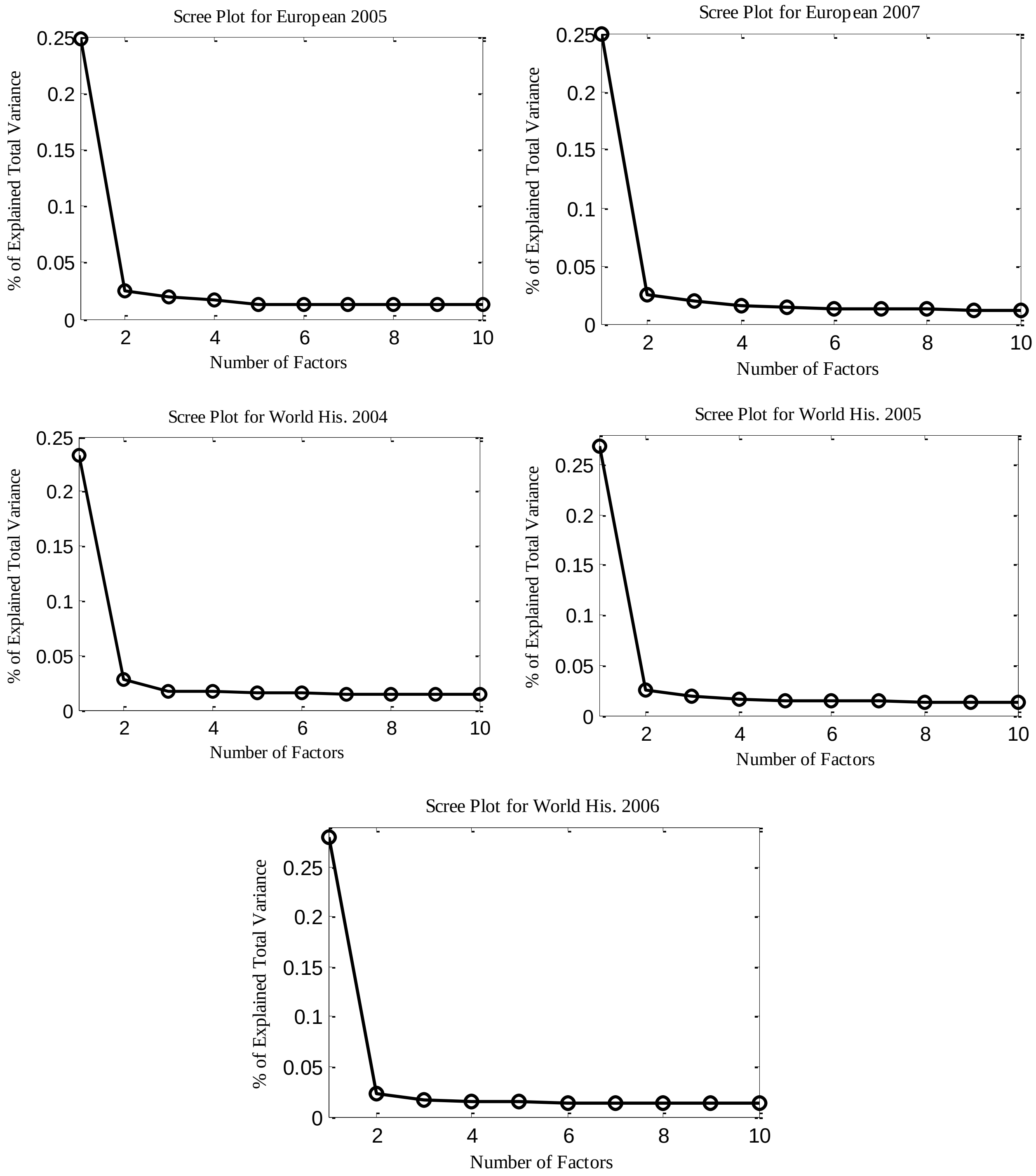


Figure 2. Scree plots for language test forms.



Scree Plot for European 2005

Scree Plot for European 2007

Scree Plot for World His. 2004

Scree Plot for World His. 2005

Scree Plot for World His. 2006

Figure 3. Scree plots for history test forms.