Center for Advanced Studies in
Measurement and Assessment

CASMA Monograph

Number 2.1

Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 1)

Michael J. Kolen
Won-Chan Lee
(Editors)

December, 2011

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

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. We worked to achieve these goals with graduate students Sarah L. Hagge (now at the National Council on State Boards of Nursing), Yi He (now at ACT, Inc.), Chunyan Liu (now at ACT, Inc.), Sonya J. Powers (now at Pearson), and Wei Wang (still a graduate student at the University of Iowa). This research program led to four dissertations (Hagge, 2010; He, 2011; Liu, 2011; & Powers, 2010) and eight presentations at professional conferences (Hagge & Kolen, 2011, April; He & Kolen, 2011, April; Lee, Hagge, He, Kolen, & Wang, 2010, April; Liu & Kolen, 2010, April; Powers, 2011, April; Powers & Brennan, 2009, April; Powers, Liu, Hagge, He, & Kolen, 2010, April; and Wang, Liu, Lee, & Kolen, 2011, April). In addition, the research program led to a monograph with two volumes, which presents the research that we conducted.

This volume (Volume 1) of the monograph contains 9 chapters. Chapter 1 provides an overview and describes methodology that is common across many of the chapters. In addition, it highlights some of the major methodological issues encountered and highlights some of the major findings.

Chapters 2 through 7 are organized around the types of criteria that are used to evaluate equating. Both item response theory (IRT) and equipercentile methods are studied in all of these chapters. In addition, the design used for data collection is the common-item nonequivalent groups design (CINEG), where nonequivalent groups of examinees take the test forms to be

equated, and the test forms contain a set of items in common. The CINEG design is used operationally to equate scores on the AP examinations.

In Chapter 2, scores on intact alternate forms of AP examinations are equated using operational data. The equating results for different methods are compared to one another to judge their similarity. In addition, bootstrap resampling methods are used to estimate standard errors of equating, which are also compared across methods.

In Chapter 3, the population dependence of equating relationships is investigated using data from intact operational forms. Groups of examinees taking the old and new forms are sampled from the total groups based on a reduced fee indicator that is related to total test score. Pseudo-groups of old and new form examinees are formed that differ by various effect size amounts in mean performance. The population dependence of the equating relationships is evaluated by comparing the equating relationships for groups with effect size differences greater than zero to the equating relationship for groups that do not differ.

In an effort to provide a more adequate criterion equating, pseudo-test forms are created in Chapters 4 through 6 as well as in Chapters 8 and 9. The studies in Chapters 4 through 6 begin with a single test form. Two pseudo-test forms are created from this single form by dividing the single form into two shorter forms with items that are in common. Because the same examinees take both the old and the new pseudo-test forms, a single group equipercentile equating can serve as the criterion equating. Various pseudo-groups that differ in average proficiency can be created and used to research various questions in CINEG equating using the single group equipercentile equating relationship as a criterion.

In Chapter 4, equating accuracy is compared for different equating methods. In Chapter 5, the influence of (a) the difference in average proficiency between the old and new form groups, (b) relative difficulty of the MC and FR items, and (c) the format representativeness of the common item set on equating accuracy is compared. In Chapter 6, the influence on equating accuracy of differences in average proficiency between the old and new groups is compared and degree of accuracy is related to the extent to which equating assumptions for various methods are violated.

In Chapter 7, using intact test forms equating methods are compared on the extent to which they achieve the equity and same distributions properties of equating.

In Chapters 8 and 9, focus is on choosing smoothing parameters in equipercentile equating using the random groups design for both log-linear presmoothing and cubic spline postsmoothing methods. The criterion equating is a single group equipercentile equating using pseudo-test forms. In Chapter 8, the accuracy of equating is investigated using fixed smoothing parameters. In Chapter 9, the accuracy of equating using various statistical approaches to choosing smoothing parameters is studied.

Various equatings conducted in the studies reported in this monograph were implemented using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009), which provides a set of open-source functions to perform all types of equating discussed by Kolen and Brennan (2004). We would like to recognize that much of the success in our research is greatly due to the availability of *Equating Recipes*.

We thank graduate students Sarah Hagge, Yi He, Chunyan Liu, Sonya Powers, and Wei Wang for their extremely hard work. We gave them considerable responsibility and freedom, and they responded admirably. We are sure that these experiences will serve them well during what we are sure will be productive careers in the field of educational measurement. Thanks to Wei Wang for her work in assembling and organizing this monograph.

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, thanks to College of Education Deans Sandra Damico and Margaret Crocco. We would especially like to thank 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, 2011
Iowa City, Iowa

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups.* Unpublished Doctoral Dissertation. The University of Iowa.

Hagge, S. L., & Kolen, M. J. (2011, April). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

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

He, Y., & Kolen, M. J. (2011, April). *A comparison of mixed-format equating methods using equating properties*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Lee, W., Hagge, S. L., He, Y., Kolen, M. J., & Wang, W. (2010, April). *Equating mixed-format tests using dichotomous anchor items*. Paper presented in the structured poster session Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple-Choice and Constructed-Response Items at the Annual Meeting of the American Educational Research Association, Denver, CO.

Liu, C. (2011). *A comparison of statistics for selecting smoothing parameters for log-linear presmoothing and cubic spline postsmoothing under a random groups design.* Unpublished Doctoral Dissertation. The University of Iowa.

Liu, C., & Kolen, M. J. (2010, April). *A comparison among IRT equating methods and traditional equating methods for mixed format tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Powers, S. J. (2010). *Impact of assumptions and group differences on equating accuracy using matched samples equating methods*. Unpublished Doctoral Dissertation. The University of Iowa.

Powers, S. J. (2011, April). *Impact of group differences on equating accuracy and the adequacy of equating assumptions*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Powers, S. J., & Brennan, R. L. (2009, April). *Multivariate generalizability analyses of mixed-format exams*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Powers, S. J., Liu, C., Hagge, S. L., He, Y., & Kolen, M. J. (2010, April). *A comparison of nonlinear equating methods for mixed-format exam*s. Paper presented in the structured poster session Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple-Choice and Constructed-Response Items at the Annual Meeting of the American Educational Research Association, Denver, CO.

Wang, W., Liu, C., Lee, W., & Kolen, M. J. (2011, April). *An investigation of item fit for large-scale mixed format tests*. Poster presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Contents

Chapter 1: Introduction and Overview[1]

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

[1] We thank Robert L. Brennan, Sarah L. Hagge, Yi He, Chunyan Liu, Sonya J. Powers, and Wei Wang for their comments on an earlier draft of this Chapter.

Abstract

This chapter provides an overview and describes methodology that is common across many chapters of this monograph. In addition, it highlights some of the major methodological issues encountered and highlights some of the major findings. It begins with a description of the AP Examinations and data followed by a brief description of equating methods and scores. The research questions and research designs used in this monograph are discussed followed by highlights of the results, discussion, and conclusions. In the discussion section, suggestions for further research are made, implications for equating practice are considered, and the importance of using an equating research design that is appropriate for the research questions is discussed.

Introduction and Overview

The research described in this monograph was conducted over a three-year period. Although we had some general goals in mind when we began, the research questions that we addressed and that are reported here evolved over time. This chapter provides an overview and describes methodology that is common across many chapters of this monograph. In addition, it highlights some of the major methodological issues encountered and some of the major findings.

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.

The chapter begins with a description of the AP Examinations and data followed by a brief description of equating methods and scores. The research questions and research designs used in this monograph are provided in the next section followed by a section highlighting the results. The chapter concludes with a discussion and conclusions section.

AP Examinations and Data

All of the studies in this monograph used data from the College Board AP Examinations. Each of these examinations contained both dichotomously scored multiple choice (MC) and polytomously scored free response (FR) items. There are more than 30 different AP examinations. We received data sets from the College Board that contained examinee item level data. Each of the data sets contained 20,000 or fewer examinee records on test forms that were administered between 2004 and 2007. In addition, we received one data set for the AP English Language examination that was administered in 2008 and that contained over 300,000 records. The studies in this monograph used subsets of these data.

Prior to 2011, MC items on AP examinations were administered and scored using a correction for guessing (also referred to as formula scoring). Formula scoring led to a considerable number of omitted responses to MC items. In 2011, the MC items on AP examinations were administered and scored using number-correct scoring. Number-correct scoring will be used in subsequent years as well.

For the research conducted in this monograph to be applicable to AP examinations in 2011 and beyond, and to most other mixed-format tests, the AP examinations used in this research were rescored using number-correct scoring. In research reported in all chapters, except

Chapter 7, examinees who responded to less than 80% of the MC items were eliminated from the data sets. A missing data imputation procedure (Bernaards & Sijtsma, 2000; Sijtsma & van der Ark, 2003) was used to impute responses to MC items that were omitted. Following imputation, all examinees included in the analyses had a score of either correct (1) or incorrect (0) for all MC items. For the study in Chapter 7, examinees with one or more omitted responses to the MC items were eliminated from the data sets.

Equating Methods

Various equating methods and data collection designs were used in the studies in this monograph, and all were described in Kolen and Brennan (2004). The methods were implemented using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009), POLYEQUATE (Kolen, 2004), or STUIRT (Kim & Kolen, 2004) software. PARSCALE (Muraki & Bock, 2003) or MULTILOG (Thissen, Chen, & Bock, 2003) software was used for IRT parameter estimation.

Operationally, the AP examinations were equated using data collected using the common-item nonequivalent groups (CINEG) design (Kolen & Brennan, 2004), which is also sometimes referred to as the nonequivalent groups with anchor test (NEAT) design (Holland & Dorans, 2006). In this design, a new form has a set of items in common with an old form. For the AP examinations, only MC items are common between forms.

Equipercentile equating methods used in this monograph for the CINEG design included the chained equipercentile (CE) method and the frequency estimation equipercentile (FE) method. Smoothing methods used with these equipercentile equating methods included the cubic spline postsmoothing method and the log-linear presmoothing method. Item response theory (IRT) true score and IRT observed score equating methods were also used. The three-parameter logistic model was used for all MC items. For the FR items, the graded response model (GRM) was used in some of the chapters and the generalized partial credit model (GPCM) was used in other chapters. All of these equating methods are described in Kolen and Brennan (2004).

In addition to the equating methods used with the CINEG design, equating methods were used with the single group design to develop criterion equatings in Chapters 4, 5, and 6. The random groups design was used in Chapters 8 and 9 to compare smoothed equipercentile equating results using cubic spline postsmoothing and log-linear presmoothing methods.

Scores

The MC items in this monograph were scored number-correct, and the raw score over the MC items was the total number of items answered correctly. The FR items had varying numbers of integer score points, depending on the examination. For operational AP examinations, the total raw score is a weighted summed score over all of the items, where the weights are typically non-integers. To simplify implementation of many of the equating methods examined in this monograph, total raw scores were calculated as summed scores using integer weights for MC and FR in all of the chapters.

Raw scores were transformed to two different types of scale scores. In the AP program, scores reported to examinees as 5, 4, 3, 2, and 1 are referred to as AP grades, and can be thought of as scale scores with 5 possible points. AP grades were used in Chapters 2 and 6. For the purpose of the research in this monograph and to use scale scores that are more similar to scale scores used in other testing programs, in Chapter 2, scale scores were formed by normalizing the raw scores (Kolen & Brennan, 2004, pp. 338-339). After comparing a variety of normalizing constants, it was found that transforming the scores to have a mean of 35 and a standard deviation of 10, and rounding and truncating the scores so that the resulting scale scores were integers in the range of 0-70, resulted in scale scores that were nearly normally distributed, with few gaps in the raw-to-scale score conversions.

Research Questions and Designs

The chapters in this monograph differ in the research questions addressed and the designs used to address them. Each of Chapters 2 through 6 addressed at least one or more of the following characteristics of CINEG equating: (a) examinee group dependence of equating relationships, (b) random equating error, and (c) systematic equating error. Because there is overlap in the questions addressed, these chapters are considered together in this section. Chapter 7 is the only chapter that addressed the first-order and second-order equity properties of equating, so it is discussed by itself. Chapters 8 and 9 examined smoothing in random groups equipercentile equating, and are discussed together.

Chapters 2-6

Highlights of the research questions addressed in Chapters 2 through 6 are provided in Table 1. The research questions focus on examining random and systematic equating error (Kolen & Brennan, 2004, pp. 23-24) and group dependence of equating relationships (Kolen &

Brennan, 2004, p. 13) for CINEG equating. Ideally, random and systematic equating errors are small and there is little group dependence of equating relationships.

The research questions listed in Table 1 tend to become more elaborate in the later chapters. Chapter 2 addresses the extent to which equating relationships and amount of random equating error differ for various equating methods. Chapter 3 focuses on group dependence of equating relationships. Chapter 4 not only considers random equating error, but also systematic equating error. Chapter 5 addresses most of the research questions that were covered in the first four chapters and also considers how much the relative difficulty of MC and FR items affects random and systematic equating error as well as how the composition of the common item sets affects such error. Chapter 6, in addition to examining group dependence of equating relationships as was done in Chapters 2 and 5, considers the extent to which such group dependence is a result of violations of statistical assumptions of various equating methods.

Highlights of design characteristics for the research presented in Chapters 2 through 6 are presented in Table 2. In general, as the research questions become more complicated, the research designs become more complicated as well.

As indicated in the first row and column of Table 2, Chapter 2 used intact examinee groups. The use of intact groups means that the groups taking the old and new forms were representative of the entire examinee groups who took the forms operationally. Chapter 2 also used intact test forms (see third row), meaning that the old and new forms were the forms given operationally. With this type of design, the similarity of equating relationships across methods can be addressed. In addition, resampling methods such as the bootstrap method can be used to estimate random equating error (Kolen & Brennan, 2004, pp. 235-245). Thus, the two research questions presented in Table 1 can be addressed using intact test forms and intact groups. However, with this design, there is no equating criterion that can be used to address the extent to which there is systematic error in equating.

Pseudo groups were used in Chapter 3. In this monograph, pseudo groups of examinees who were administered the old test form were created by sampling examinees based on a demographic variable that was external to the test. The same was done for examinees who were administered the new form of the test. A reduced fee indicator was used to form pseudo groups in Chapter 3. Examinees at different levels of the reduced fee indicator were sampled in different proportions for the old and new forms to produce examinee groups that differed in average

common item performance by a specified amount. The use of pseudo groups in Chapter 3 allowed for the investigation of dependence of equating relationships on examinee group differences.

Pseudo-test forms (von Davier, Holland, Livingston, Casabianca, Grant, & Martin, 2006) were used along with pseudo groups in Chapter 4, as indicated in Table 2. To construct pseudo-test forms, a single operational test form is divided into two shorter test forms with common items. Data from examinees who took the single operational test form are used to equate scores on the two pseudo-test forms using a single group design, since each examinee took both of the pseudo-test forms. The single group equipercentile equating of the pseudo-test forms is used as a criterion equating. The single group equipercentile equating serves as a criterion equating relative to CINEG equating because it does not involve two potential sources of error in CINEG equating: group differences and equating through a common-item set. Pseudo groups of examinee are selected from the examinees who took the operational test forms. The average common item score difference between the pseudo groups taking the old and new forms can be varied when forming the pseudo groups. CINEG equating methods are used to equate the pseudo-test forms. These CINEG equatings are compared to the criterion single group equating. This comparison is used, along with resampling procedures, as a basis for estimating systematic error in CINEG equating. In this way, the research question investigated in Chapter 4 was addressed (see Table 1). As can be seen in Table 2, a gender variable was used in Chapter 4 to form the pseudo groups, and only MC items were used as common items.

Note that with pseudo-test forms, the single group equipercentile equating is used as the criterion equating. For a CINEG equating method at a score point on the new form, systematic error (squared bias) is calculated, over samples taken in the resampling process, as the square of the difference between the mean of CINEG equated score and the criterion equated score. Random equating error (squared standard error) is the variance of the CINEG equated score over the samples taken in the resampling process. Total error (mean squared error) at a score point on the new form is calculated as the sum of systematic error and random equating error.

Pseudo groups and pseudo-test forms were used in Chapter 5 so that a single group equating could be used as the criterion for comparing equating results. As indicated in Table 1, this study examined the dependence of random and systematic equating errors on examinee group differences, the relative difficulty of MC and FR items, and the composition of the

common item set (MC only vs. MC and FR). Pseudo groups were formed using the bivariate distribution of ethnicity and parental education. Resampling procedures were used as well.

Chapter 6 also used pseudo groups and pseudo-test forms to investigate group dependence of equating. This study investigated the extent to which differences in equating relationships were due to violations of statistical assumptions in the equating methods used with CINEG design.

Chapter 7

Chapter 7 examined aspects of Lord's (1980) equity property as well as the same distributions property (Kolen & Brennan, 2004). In particular, under first-order equity, each examinee is expected to earn the same equated score on the old and new forms; under second-order equity, the precision of equated scores on the old and new forms for each examinee is the same (Kolen & Brennan, 2004, pp. 10-11); the same distributions property holds if the equated scores on the new form have the same distribution as the scores on the old form. To estimate first-order equity and second-order equity, an IRT model was assumed to hold. Intact groups and intact forms administered under the CINEG design were used to estimate first-order equity and second-order equity for 22 equatings. The relationships between a variety of factors and equity were studied.

Chapters 8 and 9

In these chapters, a pseudo-test form design and intact examinee groups were used to investigate random and systematic error in random groups equipercentile equating. Because a pseudo-test design was used, a single group equating relationship was available to use as the criterion equating. Cubic spline postsmoothing and log-linear presmoothing methods were investigated. Each of these methods has a parameter that controls the amount of smoothing. In both chapters, sample size was varied.

In Chapter 8, equating errors for a set of fixed smoothing parameters for presmoothing and postsmoothing methods were estimated and compared. Methods and degrees of smoothing associated with smaller equating error were considered to be preferable.

In Chapter 9, equating errors for automated methods of selecting smoothing parameters were estimated and compared. Automated methods with smaller associated equating error were considered to be preferable.

Highlights of Results

Results from the chapters in this monograph are summarized in this section. The section begins by examining differences among equating methods in equating relationships. The results for random error are highlighted followed by results for systematic error and group dependence in the CINEG design. Then results for first- and second-order equity are highlighted, followed by results for smoothing in the random groups design.

Equating Methods and Equating Relationships in the CINEG Design

In Chapter 2, the various CINEG equating methods tended to produce similar equating relationships. The intact groups taking the old and new forms in Chapter 2 were similar in their average scores on the common items. In conditions where there were larger group differences (Chapters 3, 5, and 6), larger differences in equating relationships among the equating methods were often found.

Random Error in Equating in the CINEG Design

In Chapters 2, 4, and 5, the amount of random error observed, as indexed by the standard error of equating using resampling procedures, tended to be in the following order from smallest to largest: IRT observed score equating, IRT true score equating, FE equating, and CE equating. In addition, in Chapter 2 it was found that smoothing methods reduced random error and that presmoothing led to slightly less random error than postsmoothing for the particular smoothing parameters that were chosen. In Chapter 5, there appeared to be little relationship between random error and the magnitude of group differences.

Systematic Error in Equating in the CINEG Design

In Chapter 4, less systematic error (bias) was found for IRT observed score equating than for IRT true score equating. Other comparisons of systematic error among equating methods were inconsistent.

In Chapter 5, as group differences increased, systematic error tended to increase. This increase in systematic error was least for the IRT methods and greatest for the FE method.

Also in Chapter 5, less systematic error was found when the common-item set contained both MC and FR items than when it contained MC items only. In addition, less systematic error was found when the relative difficulty of the MC and FR items was similar for the old and new groups than when it was dissimilar.

Group Dependence of Equating Relationships

In Chapter 3, the equating relationships for the FE method were found to be dependent on group differences. Such a relationship was not found for the other equating methods. In Chapter 6, all CINEG equating methods showed group dependence, with FE results most sensitive to group differences. Note that the group differences were larger in Chapter 6 than in Chapter 3, which might have contributed to the difference in findings between Chapter 3 and 6. In addition, pseudo-test forms were used in Chapter 6, which also might have made it easier to detect group dependence.

In Chapter 6, single group equating results were very similar across examinee groups. Further investigation of the assumptions underlying CINEG equating methods suggested that the extent to which the equating assumptions were violated was related to the amount of group dependence observed with the equating results.

Equity and the Same Distributions Property in the CINEG Design

The results in Chapter 7 suggested that between the IRT methods, first-order equity held better for IRT true score equating and second-order equity and the same distributions property held better for IRT observed score equating. Between the equipercentile methods, CE better preserved first-order equity. A higher correlation between scores on the MC and FR items was associated with better preservation of the equating properties for the IRT methods.

Smoothing in Equipercentile Equating in the Random Groups Design

In examining fixed smoothing parameters in Chapter 8, random error was reduced more as the amount of smoothing increased. Systematic error tended to increase as smoothing increased. For all sample sizes and for both presmoothing and postsmoothing, at least one of the smoothing parameters led to an increase in overall error compared to no smoothing. Especially at the larger sample sizes (5,000 and 10,000), smoothing increased overall equating error compared to no smoothing for at least some fixed smoothing parameters for both presmoothing and postsmoothing. Postsmoothing tended to lead to slightly less overall equating error than presmoothing.

In examining automated methods of selecting smoothing parameters in Chapter 9, for presmoothing the Akaike information criterion (AIC; Akaike, 1981) tended to lead to less overall equating error than the other criteria examined at most sample sizes. The ± 1 SEE (standard error of equating) criterion described in Chapter 9 was found to perform well with postsmoothing. At a

sample size of 10,000 none of the automated criteria consistently reduced overall error compared to unsmoothed equating.

Discussion and Conclusions

This section discusses many of the results, and suggestions are made for further research. Random error is discussed first followed by discussion of systematic error, choice of equating methods in practice, group dependence, first- and second-order equity, and smoothing in random groups equating. The section concludes with a discussion of designs for conducting research on equating methods.

Random Error in Equating in the CINEG Design

Random error was less for the IRT methods than for the smoothed equipercentile methods for the CINEG equatings in Chapters 2, 4, and 5. A single fixed smoothing parameter was used with the equipercentile methods in these chapters. Follow-up research should address whether the choice of other smoothing parameters might have led to different conclusions. In addition, less random error was associated with the FE method than with the CE method.

Systematic Error in Equating in the CINEG Design

Systematic error in IRT equating was found to be less than that for the equipercentile methods in Chapters 4 and 5 in some cases. In addition, with large group differences, there was more systematic error with the FE method than with the other methods.

Choice of Equating Methods in Practice

The results for systematic and random error comparisons among equating methods suggest that various CINEG equating methods have their own strengths and weaknesses. IRT methods might be preferred when the IRT assumptions hold well because they often were found to produce the smallest amount of systematic and random error. The CE method might be preferred over the FE method when large group differences exist, because the CE methods were found to produce less systematic error in this situation. The FE method might be preferred when there are small group differences, because the FE method had less random error than the CE method in this situation. Based on these findings, operational CINEG equating systems should include the two IRT methods and the two equipercentile methods when the equating procedures are intended to be applicable in situations where there might be large or small group differences or when the IRT model might or might not hold.

Group Dependence of Equating Relationships

For the CINEG design, the FE method appears to be the method that is most sensitive to group differences of the methods studied. The use of pseudo-test forms in Chapter 6 in combination with the use of larger group differences appeared to allow for more fidelity in assessing the sensitivity of CINEG equating methods to group differences in common item performance.

In Chapter 6, an attempt was made to isolate the effects of group dependence due to differences in single group equatings within each group from group dependence that is due to the violations of statistical assumptions. It appeared that the group dependence identified in Chapter 6 might have been due primarily to violations of statistical assumptions.

First- and Second-Order Equity in the CINEG Design

In Chapter 7, first- and second-order equity were evaluated by assuming that an IRT model fit the data. In further research it would be interesting to assess these equity criteria using other psychometric models such as strong true score models. Some guidelines need to be developed in terms of how to determine an acceptable level of equity for practical purposes.

Smoothing in Equipercentile Equating in the Random Groups Design

In Chapter 8, all the fixed smoothing parameters were found to reduce random error compared to no smoothing for the random groups design. In Chapter 9, the AIC criterion for presmoothing and the ±1 SEE criterion for postsmoothing appeared to reduce overall error for most sample sizes. However, for sample sizes of 5,000 and 10,000, all smoothing procedures sometimes resulted in more overall error than no smoothing. More research is needed, especially with large sample sizes. In addition, more research is needed to provide guidance about the sample sizes for which smoothing should be used. Finally, further research should involve a systematic investigation of the effects of smoothing on equating under the CINEG design.

Designs for Equating Research

In this monograph, various research designs were used to address different types of research questions. The use of intact examinee groups and intact test forms closely resembles operational testing and equating because operational test forms are equated and operational examinee groups are used to conduct equating. The research questions that can be addressed using intact examinee groups and intact test forms are restricted to questions about the similarity of the equating relationships for different equating methods and, by using resampling methods,

the amount of random error for different equating methods. The study in Chapter 2 illustrates the use of intact test forms and intact groups.

The use of intact test forms and pseudo groups can be used to relate the magnitude of group differences to the group dependence of equating relationships. The study in Chapter 3 illustrates the use of intact test forms and pseudo groups.

Pseudo-test forms provide a single group criterion equating that can be used to assess systematic error in equating. In Chapters 8 and 9, pseudo-test forms provided a criterion for comparing smoothing methods for use with the random groups design. In Chapters 4, 5, and 6, pseudo-test forms were used in combination with pseudo groups to address, for example, the following issues with the CINEG design: (a) the extent to which systematic error is related to group differences for various equating methods; (b) whether equating with MC only common items leads to more systematic equating error than equating with MC and FR common items; and (c) the relationship between group dependent equating results and violations of statistical assumptions. The single group criterion equating used in these studies was made possible by the use of pseudo-test forms. The drawback of using pseudo-test forms, compared to intact test forms, is that the pseudo-test forms are not the forms that are operationally equated or administered, so they might not represent equating as conducted in practice.

Simulation studies often are used in equating research. For example, an IRT model might be assumed, item response data generated from the IRT model, and group differences varied. A strength of such simulations is that they can be conducted without the use of actual test data and many variables can be manipulated. The drawback of simulations is that it can be difficult to make them realistic.

From a research perspective, the design used should be tailored to address the research questions of interest. To the extent that the research questions can be addressed well with intact groups and intact forms, the design should make use of intact groups and intact forms because this situation is most realistic. Pseudo groups and pseudo-test forms can be used when the research questions cannot be addressed adequately with intact forms and groups. In other situations, it might be necessary to use simulation studies so that all of the variables of interest can be manipulated in ways that cannot be done with real test data.

References

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

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.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006, March). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data* (Research Report RR-06-02). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education and Praeger.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Kolen, M. J. (2004). POLYEQUATE [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

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

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

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.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG (Version 7.03) [Computer software]. Lincolnwood, IL: Scientific Software International.

Table 1

Highlights of Research Questions for Chapters 2 through 6

Chapter	Research Questions
2	(a) How much do equating relationships differ for various equating methods? (b) How much does the amount of random equating error differ for various equating methods?
3	How much does the dependence of equating relationships on group differences differ for various equating methods?
4	How much does the amount of random and systematic error in equating differ for various equating methods?
5	(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 is the magnitude of random and systematic equating errors differentially related to the relative difficulty of MC and FR items for various equating methods? (c) To what extent is the magnitude of random and systematic equating errors differentially related to the composition of the common item set (MC only vs. MC and FR) for various equating methods?
6	(a) How much does the dependence of equating relationships on group differences differ for various equating methods? (b) To what extent is the dependence of CINEG equating relationships on group differences related to violations of statistical assumptions for the various equating methods used with the CINEG design?

Table 2

Highlights of Design Characteristics for Studies in Chapters 2 through 6

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

Chapter 2: A Comparison of IRT and Traditional Equipercentile Methods in Mixed-Format Equating

Sarah L. Hagge, Chunyan Liu, Yi He, Sonya J. Powers, Wei Wang, and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

This chapter compares four equating methods under the common-item nonequivalent groups design for mixed-format tests: the frequency estimation (FE) equipercentile method, the chained equipercentile (CE) method, the item response theory (IRT) true score equating method, and the IRT observed score equating method. In addition, presmoothing and postsmoothing procedures are used for the equipercentile methods. Data analyses are conducted using real data from three Advanced Placement (AP) Examinations. Overall, the results show that different equating methods produced similar equating functions, moments, and grade distributions. Additionally, average standard errors of equating were smaller for IRT observed equating than for the IRT true score method, for FE than for CE, for the smoothed than for the unsmoothed methods, and for the presmoothed than for the postsmoothed method. Meanwhile, the IRT methods produced smaller average standard errors of equating compared to the equipercentile methods for most of the equating relationships investigated in this study.

A Comparison of IRT and Traditional Equipercentile Methods in Mixed-Format Equating

This chapter compares the results from four non-linear equating methods in mixed-format test equating: the frequency estimation equipercentile (FE) method, the chained equipercentile (CE) method, the item response theory (IRT) true score equating method, and the IRT observed score equating method. The first two are commonly-used traditional equipercentile methods, and the latter two are IRT methods. In addition, presmoothing and postsmoothing procedures are used for the equipercentile methods. Results are evaluated in terms of equated score distributions and moments, bootstrap standard errors of equating, and grade distributions.

For the two equipercentile methods, von Davier, Holland, and Thayer (2004) showed mathematically that the FE and CE methods produce equivalent results when the groups taking the old and new form are equivalent or when the common-item scores and total test scores are perfectly correlated. However, these situations typically do not exist in practice, so the two methods often produce different results (Braun & Holland, 1982; Harris & Kolen, 1990). Kolen and Brennan (2004, p. 146) argued that the FE method might be theoretically preferable to the CE method, because the FE method defines the equipercentile relationship for a specified examinee group, referred to as the synthetic population, whereas the CE method does not have a clear definition of the examinee group. In addition, the CE method often involves equating tests with significantly different lengths (the total test versus the common item set). Previous research indicates that the FE method produces smaller standard errors of equating than the CE method. However, the CE method produces less biased results than the FE method when there is a large difference in ability between the groups of examinees taking the forms to be equated (Holland, Sinharay, von Davier, & Han, 2008; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008). On the other hand, Ricker and von Davier (2007) found that when the common-item set was relatively short compared to the total test, equating results obtained by the FE method had less bias than the results from the CE method. Based on this research, it appears that the FE method might be preferable when the group differences are small and the CE method preferable when there are larger group differences.

When comparing equating results of IRT true score and IRT observed score equating methods, Kolen (1981) and Han, Kolen, and Pohlmann (1997) found that the two methods produced somewhat different results under the random groups design. However, a study by Lord

and Wingersky (1984) showed that under the common-item nonequivalent groups design, the two methods led to very similar results. In addition, when using equating a test to itself as the criterion, Lord and Wingersky (1984) found that both methods produced very close equating relationships to the identity equating relationship, and Han et al. (1997) found that equating results obtained using the IRT true score method were closer to the criterion than those obtained using the IRT observed score method. Furthermore, Tsai, Hanson, Kolen, and Forsyth (2001) found that standard errors of equating were slightly smaller with the IRT observed score method than with the IRT true score method for a test containing dichotomously-scored items, and Cho (2008) had similar findings for a test containing polytomously-scored items. However, Cho (2008) also found that when sample size was very small, such as 250 students per form, the IRT true score method yielded smaller standard errors of equating than did the IRT observed score method.

Comparing the traditional equipercentile method with the IRT methods under the random groups design, a study by Kolen (1981) found that the IRT methods with the three-parameter logistic (3PL) model yielded more consistent results than the unsmoothed equipercentile method, and Han et al. (1997) found that IRT methods yielded results that were closer to the criterion than the equipercentile method when equating a test to itself. However, for small sample sizes, the equipercentile method produced smaller standard errors of equating than IRT methods (Cho, 2008).

In sum, the existing literature shows that different equating methods lead to different equating functions, although the extent of differences in equated scores across methods might vary from situation to situation. In addition, the FE method leads to lower standard errors of equating than the CE method, and the IRT observed score method leads to lower standard errors of equating than the IRT true score method when sample sizes are large. However, all of the studies just reviewed were based on single format tests. It is unclear how well these results will generalize to mixed-format test equating. The purpose of this chapter was to compare results for these nonlinear equating methods for tests with both dichotomously-scored multiple choice (MC) items and polytomously-scored free response (FR) items.

Method

Data

Equating and scaling was conducted for five equating relationships using three AP Exams:

Biology, English Language and Composition (referred to here as English Language), and French Language. These exams included at least one MC section and at least one FR section. The administration times, numbers of items, section weights, and composite score scales varied across the exams. The equating links and original sample sizes for each form are provided in Table 1. Specifically, the data for Biology 2004-2006, Biology 2005-2007, English Language 2004-2007, French Language 2004-2006, and French Language 2005-2007 were used. Because the data in this study were collected when AP Exams were administered using formula scoring, and examinees were informed that a fraction of a point would be deducted if an MC item was answered incorrectly, this led to a substantial amount of missing responses for MC items. Across the exams, approximately 10–30% of examinees responded to all items. In this study, number-correct scoring was used. Number-correct scoring was approximated by imputing missing data for omitted items using a two-way procedure described by Bernaards and Sijtsma (2000) and Sijtsma and van der Ark (2003). To reduce the amount of missing data when calculating number-correct scores, examinees who responded to less than 80% of MC items were eliminated prior to the imputation process. Table 1 also provides the actual sample sizes after imputation that were used for the current study.

Section Weights

Operationally, the MC and FR sections are weighted by non-integer values to ensure that their contribution to the composite, in terms of proportion of points, meets test development requirements. One benefit of non-integer weights is that the same composite score range can be ensured even if the number of items or the number of score categories within an item changes from year-to-year. However, use of non-integer weights results in non-integer scores for examinees, whereas many psychometric procedures and programs are designed for integer scores. To avoid rounding non-integer scores, integer weights were considered in the current study. Integer weights were selected so that the MC and FR contributions to the composite score were similar to their operational contributions. Integer weights used in the study and the resultant score ranges are provided in Table 2. As can be seen from Table 2, one drawback of using integer weights is that they result in different composite score ranges for the two forms that are equated. For each form, raw score means and standard deviations were calculated for the composite score and the common-item score, which are listed in Table 3. The covariances and correlations between the composite scores and the common-item scores are also shown in Table 3. For each equating

relationship, common-item effect size between groups taking the old and new forms was less than .10, indicating that the two groups were very similar.

Score Scale for the Old Form

To use scale scores that are similar to scale scores used in other testing programs, scale scores for the base form were found by normalizing the weighted composite score using procedures described in Kolen and Brennan (2004, pp. 338-339). In this procedure, percentile ranks were found for the weighted composite score, an inverse normal transformation was applied to the percentile ranks to produce *z*-scores that were approximately normal with mean of 0 and standard deviation of 1, and these *z*-scores were linearly transformed to have a mean of approximately 35 and standard deviation of approximately 10. Finally, the scale scores were rounded to integers and truncated to be between 0 and 70.

Equating Methods

Two traditional methods and two IRT methods were used for equating: FE, CE, IRT true score equating, and IRT observed score equating. For the FE and IRT observed score equating methods, the synthetic weight of 1 was assigned to the group of examinees taking the new form. Bivariate log-linear presmoothing and cubic spline postsmoothing procedures were used with the equipercentile methods. For presmoothing, the composite scores were modeled using a polynomial of degree 6, the common item scores with a polynomial of degree 6, and one cross-product was included. The postsmoothing parameter used in the analyses was 0.1, which was the highest value that resulted in a smoothed equating relationship within error bands of plus and minus one raw score standard error within score range percentile ranks of 0.5 and 99.5. Beyond this range, the equating function was estimated using linear interpolation due to the small number of examinees scoring at these score points. For the traditional methods, equating and smoothing were implemented using *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009).

For IRT equating, item parameters were estimated using MULTILOG (Thissen, Chen, & Bock, 2003). The 3PL model and the graded response model (GRM) were assumed to estimate item parameters for the MC and FR items respectively. MULTILOG default priors were used for the item estimation process. Because MULTILOG does not provide an estimate of the posterior ability distribution, which is necessary for IRT observed score equating, quadrature weights from a standard normal distribution were used. After obtaining item parameter estimates and

quadrature weights, STUIRT (Kim & Kolen, 2004) was used to transform the new form item parameters and ability distribution to the old form scale using the Haebara method (Haebara, 1980) for scale transformation. Both IRT true and observed score equating were conducted using PolyEQUATE (Kolen, 2004).

Evaluation Criteria

Results were evaluated using three criteria: equated score distributions and moments, bootstrap standard errors of equating, and grade distributions. Equated score distributions and the first four moments could be directly obtained from the *ER* output for the traditional methods and from the PolyEQUATE output for the IRT methods.

Standard errors of equating index the amount of random error in equating relationships and estimate the degree of inaccuracy of equated scores. Kolen and Brennan (2004) defined the standard error of equating as follows: First, let x_i be a particular score on Form X (new form), define $e\hat{q}_Y(x_i)$ as its estimated equivalent on Form Y (old form), and define $E[e\hat{q}_Y(x_i)]$ as the expected equivalent, where *E* is the expectation over random samples from the population; second, for a given equating, at score x_i, equating error is defined as the difference between the estimated equivalent and the expected equivalent, i.e., $e\hat{q}_Y(x_i) - E[e\hat{q}_Y(x_i)]$; finally, the variance of equating error is defined as the expectation of squared differences of $e\hat{q}_Y(x_i) - E[e\hat{q}_Y(x_i)]$ over replications, and the standard error of equating is the square root of the equating error variance, i.e.,

$$se[e\hat{q}_Y(x_i)] = \sqrt{E\{e\hat{q}_Y(x_i) - E[e\hat{q}_Y(x_i)]\}^2} \,. \qquad (1)$$

In the present study, to estimate standard errors of equating, the bootstrap method (Efron, 1982; Efron & Tibshirani, 1993) was used. The general procedure of the bootstrap method for estimating standard errors of equating when equating Form X with a sample size of N_X to Form Y with a sample size of N_Y is:

1. Draw a random bootstrap sample with replacement of size N_X from the sample of N_X examinees who took Form X.
2. Draw a random bootstrap sample with replacement of size N_Y from the sample of N_Y examinees who took Form Y.
3. With the bootstrap samples drawn in steps 1 and 2, conduct an equating using a specified equating method.

4. Repeat steps 1 through 3 *R* times. At each integer score x_i of Form X, refer to its estimated equivalent in Form Y as $\hat{e}_{Yr}(x_i)$ for the r-th replication.

5. The standard error of equating is the standard deviation of the *R* estimates at each integer score x_i. Mathematically, it is estimated by

$$s\hat{e}_{boot}[\hat{e}_Y(x_i)] = \sqrt{\frac{\sum_r [\hat{e}_{Yr}(x_i) - \hat{e}_{Y\bullet}(x_i)]^2}{R-1}}, \quad (2)$$

where

$$\hat{e}_{Y\bullet}(x_i) = \frac{\sum_r \hat{e}_{Yr}(x_i)}{R}. \quad (3)$$

To estimate standard errors of equating for the FE and CE methods, *ER* was used with 1,000 bootstrap replications. For the IRT methods, free software R (R Development Core Team, 2010) was used to call MULTILOG, STUIRT, and PolyEQUATE, and provide 1,000 bootstrap replications.

The third criterion was the distribution of AP grades. Operationally, each examinee receives an AP grade, from 1 to 5, based on his or her composite score. However, the cut scores used to convert the operational composite scores to grades could not be directly used in the current study. Those cut scores were based on the original data with formula scoring, whereas in the current study, imputation, number-correct scoring, and integer weighting were applied to the operational data. In order to obtain the percentage of examinees earning each grade, cut scores needed to be approximated. This process was done using an equipercentile method where approximately the same proportion of examinees received a score of 1, of 2 or less, of 3 or less, and of 4 or less, in the study data as in the original operational data. Cut scores were found for the old form of each equating link, and examinees taking the old forms were assigned a grade by comparing their composite scores against these cut scores. For those examinees taking the new forms, their grades were obtained by applying the cut scores for the old form to their equated scores. The grade distributions resulting from various equating methods were then compared with one another and with the old form grade distribution.

Results

Equated Score Distribution

Tables 4 to 8 provide the first four moments of raw score equivalents for the four equating methods as well as the three smoothing conditions. These moments are for the group

taking the new form with scores on the equated old form raw score scale. Each table is for one equating link. Across the methods and smoothing conditions, the differences between the largest mean and the smallest mean ranged from 0.26 (for Biology 2004-2006) to 1.46 (for English Language and Composition) score points for different equating relationships. Considering the large number of score points (241 for English Language and Composition and 310 or more for other exams), the differences in means across methods and smoothing conditions are small. In fact, for English Language and Composition, the difference between the largest and the smallest means across equating methods was less than 0.05 standard deviation units. Likewise, differences in the higher moments are small across equating methods.

Tables 9 to 13 provide the first four scale score moments for each of the five exams. Both rounded and unrounded scale score moments are included. Because the score ranges were compressed to be in the range of 0 to 70, the differences between the largest mean and the smallest mean were even lower: from 0.06 (for unrounded scale scores of Biology 2004-2006) to 0.53 (for rounded scale scores of English Language and Composition 2004-2007). This indicates that different equating methods produced very similar average equated scores. In addition, the first four unsmoothed moments were very closely matched by both pre and postsmoothing methods. Differences in the first four moments between FE and CE were also fairly consistent across smoothing methods.

Figures 1 to 5 present comparisons of old form equivalents across methods for each of the equating relationships. For the purpose of legibility, only the middle part of the score range, i.e., between percentile ranks of 0.5 and 99.5, was plotted. Beyond this score range, the equating function was estimated using linear interpolation due to the small number of examinees scoring at these score points. Each figure contains four plots: Plot (A) compares equivalent scores across IRT methods and traditional methods with presmoothing, Plot (B) compares equivalents across IRT methods and traditional methods with postsmoothing, Plots (C) and (D) are comparisons of unsmoothed, presmoothed, and postsmoothed equivalents for FE and CE, respectively. In Plots (A) and (B) of these figures, it is evident that IRT true and observed score methods resulted in similar equated scores, and the FE and CE methods resulted in similar equated scores. The extent to which IRT and traditional equating methods resulted in similar equated scores differed according to exam. Generally, throughout the middle of the score range, the equating relationships for IRT and traditional methods differed by approximately two to three raw score

points or less. The differences between the methods were larger at the upper and lower ends of the score range where few examinees scored. For Biology 2005-2007, for example, at the upper end of the score distributions, the equating relationships differed by more than eight raw score points. In Plots (C) and (D) of Figures 1 to 5, it can be seen that the equated score distributions with either presmoothing or postsmoothing were not as bumpy as those without smoothing. In addition, the postsmoothed equivalents appeared to follow the unsmoothed equivalents more closely than the presmoothed equivalents. This pattern was consistent across exams.

Standard Errors of Equating

Figures 6 to 10 present the conditional raw score standard errors of equating for the five equating relationships, respectively. Like Figures 1 to 5, each figure contains four plots, and each plot compares standard errors of equating across equating/smoothing methods. In all five figures, Plot (A) and Plot (B) show that that for all the five equating relationships, standard errors of equating were smaller for IRT equating methods than for traditional equating methods across the majority of the score range. Additionally, it is evident that between the two IRT methods, the IRT observed score equating resulted in slightly lower standard errors of equating than the IRT true score method; between the two traditional methods, standard errors for the FE method were lower than those for the CE method. Plots (C) and (D) in Figures 6 to 10 compare standard errors of equating across the three smoothing conditions for the FE and CE methods, respectively. For all five equating relationships and both the FE and CE methods, standard errors of equating were lower for presmoothing and postsmoothing than for the unsmoothed method across the majority of the score range. Between the two smoothing methods, presmoothing always produced smaller standard errors than did postsmoothing.

Tables 14 to 16 contain the weighted average standard errors of equating for IRT and traditional equating methods for the five equating relationships, using the new form relative frequencies as weights (Kolen & Brennan, pp. 243-244). In addition, the ratios of the standard error of equating (SE) to the standard deviation (SD) are provided to make it possible to compare values across exams and equating relationships. Values in Table 14 are for raw scores, and values in Tables 15 and 16 are for unrounded scale scores and rounded scale scores, respectively. All of the SE/SD values for a given scaling method were fairly close within and across equating methods. With the exception of English 04-07, the average standard errors of equating for IRT methods were smaller than for traditional methods for raw scores and rounded scale scores,

regardless of the smoothing method used. Between the IRT methods, average standard errors for the observed score method were smaller than those for the true score method, and between the traditional methods, FE yielded lower average standard errors than CE. Additionally, smoothing reduced standard errors of equating, and presmoothing led to lower average standard errors than postsmoothing. These patterns are consistent across the three scales (raw scores, unrounded scale scores, and rounded scale scores).

Comparing the values in Table 15 with those in Table 16, it is evident that the average standard errors of equating for unrounded scale scores were consistently lower than those for rounded scale scores, across all equating methods and all exams. This is because rounding introduces additional random error to the equating results.

Grade Distributions

Tables 17 to 21 contain the proportion of examinees receiving each grade based on the different equating and smoothing methods for each of the five equating relationships. The first column of data in each table is the proportion of examinees receiving each grade for the old form. The remaining columns are for each of the different equating and smoothing methods. The tables show that across all equating and/or smoothing methods, the proportion of examinees receiving a given grade differed by approximately 0.0225 or less, indicating that different methods resulted in either the same or similar grade distributions. In addition, for all five equating relationships, the equated new form had very similar grade distributions as the old form, which was expected given the similarity of the old and new form groups in terms of their common item performance.

Discussion and Limitations

Four equating methods (FE, CE, IRT true score, and IRT observed score) were used to estimate the equating relationship between five sets of AP tests. In addition, for the FE and CE methods, three smoothing conditions were used: no smoothing, presmoothing, and postsmoothing. IRT true and observed score equating methods tended to produce results that were similar to one another. The FE and CE methods tended to produce results that were similar to one another. However, the IRT methods tended to produce results that were somewhat different from the equipercentile methods. Overall, the differences between equating results for the different equating and smoothing methods were small. Furthermore, classification consistency of students into the five AP grades was very similar across equating and smoothing methods.

The groups taking the old and new forms were very similar in performance in the data used for this study. This may explain why the equating functions obtained with different equating methods were so similar. When group differences are larger, greater differences in equating results across methods might be observed.

In all cases smoothing led to smoother equipercentile equating relationships than unsmoothed methods, as expected. In general, the postsmoothing method led to equating relationships that were more similar to the unsmoothed equating relationships than did the presmoothed method. However, only one smoothing parameter was used for the postsmoothing method (S=.1) and only one set of smoothing parameters (6, 6, 1) was used for the presmoothing method. The finding of smoother equipercentile equivalents for presmoothing than for postsmoothing in this study might have been due to choice of smoothing parameters: The postsmoothing parameter of 0.1 was a relatively low degree of smoothing, whereas the presmoothing parameterization of 6, 6, and 1 was a relatively high degree of smoothing, and thus the distributions were smoothed to a different extent. The effects of choice of smoothing parameters on the smoothness of equating relationships for mixed-format tests is an area for further research (also see Chapters 8 and 9).

For raw scores, unrounded scale scores, and rounded scale scores, the average standard errors of equating were smaller for the IRT methods than for the equipercentile methods (except English 04-07), for IRT observed equating than for the IRT true score method, for the FE than for the CE method, for the smoothed than for the unsmoothed methods, and for the presmoothed than for the postsmoothed method.

The standard errors of equating index random error. Although methods with smaller standard errors have less random error, they might have more systematic error. Thus, methods with smaller standard errors are not necessarily preferable. The magnitude of systematic error is more difficult to assess than the magnitude of random error, often requiring simulation studies. A full comparison of methods would involve assessing both sources of error.

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.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

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.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

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

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.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

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

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

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing [computer software]. R Foundation for Statistical Computing, Vienna, Austria.

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

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.

Sinharay, S., & Holland, P. W. (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.

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

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.

Table 1

Equating Links and Sample Sizes

		Old Form		New Form	
Exam	Equating Link	Original N	Imputed N	Original N	Imputed N
Biology	2004-2006	20,000	15,075	20,000	16,899
	2005-2007	20,000	16,185	20,000	16,819
English Language	2004-2007	20,000	15,820	20,000	16,882
French Language	2004-2006	13,480	12,349	15,154	13,403
	2005-2007	14,495	13,571	15,287	13,982

Table 2

Number of Items, Maximum FR scores, Section Weights, and Composite Score Range

Exam	# of MC Items	# of FR Items	Max FR Score	MC Weight	FR Weight	Composite Score Range
Biology 2004	99	4	10 each	2	3	0 - 318
Biology 2006	98	4	10 each	2	3	0 - 316
Biology 2005	98	4	10 each	2	3	0 - 316
Biology 2007	99	4	10 each	2	3	0 - 318
English 2004	53	3	9 each	2	5	0 - 241
English 2007	52	3	9 each	2	5	0 - 239
French 2004	80	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 5, 3	0 - 310
French 2006	84	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 5, 3	0 - 318
French 2005	79	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 5, 3	0 - 318
French 2007	85	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 5, 3	0 - 320

Note. For the FR items of the French Language Exams, a weight of 1 was assigned to FR 1 and FR 2, a weight of 5 was assigned to FR 3, and a weight of 3 was assigned to FR items 4 to 8.

Table 3

Raw Score Means and Standard Deviations for Each Test and Each Set of Common Items, and Covariances and Correlations between Composite Scores and Common Item Scores

Equating Relationship	Group	Score	Mean	SD	Covariance	Correlation	Effect Size
Biology 2004-2006	1	X	179.0555	57.4094	532.8295	.9110	-.0536
	1	V	34.0928	10.1888			
	2	Y	178.0025	57.4910	536.9402	.9068	
	2	V	33.5435	10.3007			
Biology 2005-2007	1	X	164.5396	56.0686	439.6975	.8755	.0148
	1	V	30.3331	8.9581			
	2	Y	176.4515	53.3873	403.8183	.8727	
	2	V	30.4633	8.6675			
English 2004-2007	1	X	142.0768	33.1099	156.4827	.7650	.0083
	1	V	27.6483	6.1787			
	2	Y	147.7683	33.8501	161.3754	.7574	
	2	V	27.7001	6.2946			
French 2004-2006	1	X	187.4377	54.5105	499.3972	.8955	.0935
	1	V	29.1145	10.2314			
	2	Y	185.4447	54.0371	508.6514	.9004	
	2	V	30.0816	10.4547			
French 2005-2007	1	X	190.8128	53.5398	488.0502	.8953	.0544
	1	V	31.7511	10.1820			
	2	Y	182.7078	50.1573	466.1852	.9001	
	2	V	32.3086	10.3265			

Table 4

Biology 2004-2006 Equated Raw Score Moments

	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	FE	CE	FE	CE	FE	CE	True	Observed
Mean	180.8263	181.0827	180.8352	181.0804	180.8278	181.0799	180.9858	180.9695
SD	57.0056	57.0001	57.0186	57.0118	56.9881	56.9816	56.4846	56.7485
Skew	-.3248	-.3375	-.3253	-.3384	-.3251	-.3379	-.3776	-.3741
Kurt	2.4267	2.4385	2.4286	2.4408	2.4233	2.4343	2.4453	2.4503

Table 5

Biology 2005-2007 Equated Raw Score Moments

	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	FE	CE	FE	CE	FE	CE	True	Observed
Mean	175.7900	175.8797	175.7681	175.8842	175.7774	175.8684	176.2379	176.2921
SD	54.6808	55.1959	54.6704	55.1929	54.6806	55.1777	56.4290	56.5089
Skew	-.3022	-.2879	-.3031	-.2878	-.3040	-.2897	-.1707	-.1668
Kurt	2.4582	2.4295	2.4579	2.4275	2.4567	2.4257	2.3377	2.3353

Table 6

English Language 2004-2007 Equated Raw Score Moments

	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	FE	CE	FE	CE	FE	CE	True	Observed
Mean	147.4885	147.3390	147.5375	147.3363	147.5066	147.3436	146.0762	146.1751
SD	33.5537	33.3101	33.5393	33.3178	33.5173	33.2806	32.7738	33.2493
Skew	-.4326	-.4190	-.4329	-.4208	-.4310	-.4143	-.4812	-.4725
Kurt	2.9690	2.9836	2.9711	2.9862	2.9575	2.9558	3.0308	3.0733

Table 7

French Language 2004-2006 Equated Raw Score Moments

	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	FE	CE	FE	CE	FE	CE	True	Observed
Mean	180.9383	180.4782	181.0109	180.4757	180.9479	180.4874	179.6331	179.6558
SD	53.2380	52.7426	53.2762	52.7378	53.2601	52.7361	52.1630	52.2689
Skew	-.2368	-.2283	-.2342	-.2282	-.2335	-.2257	-.3319	-.3302
Kurt	2.3586	2.3626	2.3546	2.3640	2.3557	2.3597	2.4688	2.4751

Table 8

French Language 2005-2007 Equated Raw Score Moments

	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	FE	CE	FE	CE	FE	CE	True	Observed
Mean	180.3032	180.0596	180.3049	180.0560	180.2929	180.0560	178.9188	178.9262
SD	49.6806	49.4017	49.6382	49.3988	49.7008	49.4395	50.4540	50.4546
Skew	-.2655	-.2455	-.2650	-.2447	-.2661	-.2463	-.3317	-.3281
Kurt	2.5440	2.5509	2.5401	2.5474	2.5513	2.5611	2.5942	2.5902

Table 9

Biology 2004-2006 Equated Scale Score Moments

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Mean	Rounded	35.4745	35.5102	35.4680	35.5083	35.4549	35.5020	35.4439	35.4458
	Unrounded	35.4745	35.5107	35.4758	35.5087	35.4735	35.5084	35.4515	35.4554
SD	Rounded	9.9673	9.9654	9.9652	9.9582	9.9310	9.9410	9.7945	9.8339
	Unrounded	9.9502	9.9551	9.9501	9.9560	9.9330	9.9385	9.7740	9.8508
Skew	Rounded	-.0340	-.0604	-.0405	-.0672	-.0405	-.0655	-.1291	-.1281
	Unrounded	-.0369	-.0595	-.0363	-.0631	-.0403	-.0646	-.1326	-.1298
Kurt	Rounded	2.9995	3.0228	2.9951	3.0098	2.9746	2.9966	2.9427	2.9844
	Unrounded	3.0126	3.0301	2.9934	3.0214	2.9706	2.9928	2.9464	2.9898

Table 10

Biology 2005-2007 Equated Scale Score Moments

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Mean	Rounded	34.8998	34.9435	34.9196	34.9524	34.9212	34.9450	35.1540	35.1645
	Unrounded	34.9162	34.9573	34.9120	34.9576	34.9122	34.9533	35.1488	35.1647
SD	Rounded	10.2517	10.3729	10.2577	10.3678	10.2488	10.3596	10.7549	10.7631
	Unrounded	10.2528	10.3721	10.2476	10.3723	10.2501	10.3635	10.7404	10.7625
Skew	Rounded	.0209	.0311	.0115	.0272	.0080	.0220	.2318	.2364
	Unrounded	.0113	.0283	.0097	.0276	.0014	.0209	.2364	.2433
Kurt	Rounded	2.9030	2.8778	2.8993	2.8774	2.8912	2.8650	2.8846	2.8737
	Unrounded	2.9120	2.8859	2.9070	2.8850	2.8985	2.8675	2.8988	2.8913

Table 11

English Language 2004-2007 Equated Scale Score Moments

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Mean	Rounded	34.8847	34.8796	34.9367	34.8751	34.9168	34.8641	34.4075	34.4657
	Unrounded	34.9048	34.8492	34.9203	34.8496	34.9082	34.8490	34.4246	34.4790
SD	Rounded	9.8941	9.8431	9.9103	9.8497	9.9090	9.8373	9.5276	9.7203
	Unrounded	9.9063	9.8370	9.9041	9.8376	9.8991	9.8318	9.5415	9.7203
Skew	Rounded	.0317	.0353	.0309	.0337	.0266	.0369	-.0593	-.0297
	Unrounded	.0258	.0385	.0286	.0388	.0240	.0398	-.0603	-.0310
Kurt	Rounded	2.9962	3.0177	2.9832	2.9985	2.9719	2.9933	2.9843	3.0217
	Unrounded	2.9798	3.0103	2.9821	3.0094	2.9744	2.9933	2.9681	3.0350

Table 12

French Language 2004-2006 Equated Scale Score Moments

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Mean	Rounded	34.1383	34.0561	34.1672	34.0660	34.1542	34.0595	33.8011	33.8123
	Unrounded	34.1512	34.0567	34.1679	34.0560	34.1568	34.0608	33.8092	33.8155
SD	Rounded	9.7464	9.6058	9.7534	9.6247	9.7571	9.6380	9.4452	9.4938
	Unrounded	9.7400	9.6130	9.7501	9.6152	9.7499	9.6203	9.4386	9.4735
Skew	Rounded	.0511	.0553	.0528	.0588	.0652	.0700	-.1143	-.1154
	Unrounded	.0489	.0582	.0548	.0619	.0629	.0709	-.1111	-.1163
Kurt	Rounded	3.0510	3.0402	3.0376	3.0378	3.0765	3.0584	3.1710	3.2047
	Unrounded	3.0500	3.0425	3.0317	3.0433	3.0655	3.0714	3.1818	3.2092

Table 13

French Language 2005-2007 Equated Scale Score Moments

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Mean	Rounded	34.5047	34.4730	34.5225	34.4801	34.5263	34.4768	34.2058	34.2224
	Unrounded	34.5148	34.4685	34.5144	34.4683	34.5145	34.4692	34.2179	34.2210
SD	Rounded	9.8632	9.7993	9.8394	9.8021	9.8769	9.8082	9.9699	9.9729
	Unrounded	9.8571	9.8047	9.8444	9.8023	9.8633	9.8222	9.9726	9.9725
Skew	Rounded	.0517	.0899	.0534	.0926	.0596	.0940	-.0430	-.0375
	Unrounded	.0522	.0926	.0526	.0942	.0582	.0961	-.0447	-.0421
Kurt	Rounded	3.0265	3.0655	3.0196	3.0610	3.0473	3.1076	3.0803	3.0648
	Unrounded	3.0291	3.0596	3.0279	3.0591	3.0533	3.1052	3.0828	3.0725

Table 14

Average Raw Score Standard Errors of Equating for Traditional and IRT Equating

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Biology	SE	.6478	.7429	.4971	.5399	.5923	.6668	.4490	.4345
04-06	SE/SD	.0114	.0130	.0087	.0095	.0104	.0117	.0079	.0077
Biology	SE	.6661	.7644	.5149	.5703	.6032	.7013	.5039	.4892
05-07	SE/SD	.0122	.0138	.0094	.0103	.0110	.0127	.0089	.0087
English	SE	.5369	.6259	.4310	.5017	.4649	.5425	.4325	.3817
04-07	SE/SD	.0160	.0188	.0129	.0151	.0139	.0163	.0132	.0115
French	SE	.7244	.8403	.5697	.6360	.6653	.7647	.5308	.5141
04-06	SE/SD	.0136	.0159	.0107	.0121	.0125	.0145	.0102	.0098
French	SE	.6681	.7871	.5235	.5911	.6135	.7118	.4792	.4651
05-07	SE/SD	.0134	.0159	.0105	.0120	.0123	.0144	.0095	.0092

Table 15

Average Unrounded Scale Score Standard Errors of Equating for Traditional and IRT Equating

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Biology	SE	.1529	.1759	.1145	.1337	.1281	.1479	.0991	.0946
04-06	SE/SD	.0154	.0177	.0115	.0134	.0129	.0149	.0101	.0096
Biology	SE	.1574	.1840	.1226	.1414	.1355	.1612	.1189	.1142
05-07	SE/SD	.0154	.0177	.0120	.0136	.0132	.0156	.0111	.0106
English	SE	.1679	.1943	.1395	.1624	.1449	.1676	.1371	.1207
04-07	SE/SD	.0170	.0198	.0141	.0165	.0146	.0170	.0144	.0124
French	SE	.1730	.2003	.1337	.1529	.1481	.1739	.0963	.0942
04-06	SE/SD	.0178	.0208	.0137	.0159	.0152	.0181	.0102	.0099
French	SE	.1661	.1959	.1321	.1532	.1444	.1710	.1112	.1073
05-07	SE/SD	.0169	.0200	.0134	.0156	.0146	.0174	.0111	.0108

Table 16

Average Rounded Scale Score Standard Errors of Equating for Traditional and IRT Equating

		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
Biology	SE	.2700	.2923	.2352	.2496	.2540	.2711	.2204	.2147
04-06	SE/SD	.0271	.0293	.0236	.0251	.0256	.0273	.0225	.0218
Biology	SE	.2772	.3011	.2461	.2602	.2624	.2867	.2442	.2398
05-07	SE/SD	.0270	.0290	.0240	.0251	.0256	.0277	.0227	.0223
English	SE	.2879	.3131	.2607	.2833	.2699	.2948	.2660	.2538
04-07	SE/SD	.0291	.0318	.0263	.0288	.0272	.0300	.0279	.0261
French	SE	.2889	.3138	.2518	.2688	.2708	.2928	.2182	.2138
04-06	SE/SD	.0296	.0327	.0258	.0279	.0278	.0304	.0231	.0226
French	SE	.2829	.3102	.2477	.2644	.2661	.2906	.2358	.2323
05-07	SE/SD	.0287	.0317	.0252	.0270	.0269	.0296	.0237	.0233

Table 17

Distributions of AP Grades for Biology 2004-2006

AP Grade	Old Form Score Distribution	Equated Score Distribution							
		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
1	.1407	.1288	.1288	.1288	.1288	.1288	.1288	.1288	.1288
2	.2460	.2387	.2324	.2387	.2324	.2387	.2324	.2324	.2324
3	.2213	.2172	.2234	.2172	.2234	.2172	.2234	.2181	.2181
4	.1991	.2102	.2102	.2171	.2102	.2102	.2102	.2225	.2225
5	.1929	.2052	.2052	.1982	.2052	.2052	.2052	.1982	.1982

Table 18

Distributions of AP Grades for Biology 2005-2007

AP Grade	Old Form Score Distribution	Equated Score Distribution							
		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
1	.1522	.1626	.1626	.1626	.1672	.1626	.1626	.1758	.1758
2	.2364	.2339	.2339	.2405	.2359	.2339	.2339	.2337	.2337
3	.2318	.2247	.2175	.2109	.2109	.2247	.2175	.2045	.2045
4	.1980	.1939	.1958	.2011	.1958	.1939	.1958	.1855	.1855
5	.1816	.1849	.1901	.1849	.1901	.1849	.1901	.2005	.2005

Table 19

Distributions of AP Grades for English Language 2004-2007

AP Grade	Old Form Score Distribution	Equated Score Distribution							
		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
1	.0853	.0830	.0830	.0830	.0785	.0830	.0830	.0830	.0862
2	.3255	.3331	.3331	.3331	.3474	.3331	.3331	.3542	.3510
3	.3198	.3220	.3220	.3220	.3123	.3220	.3220	.3123	.3123
4	.1723	.1673	.1673	.1673	.1746	.1673	.1746	.1750	.1690
5	.0970	.0945	.0945	.0945	.0872	.0945	.0872	.0755	.0814

Table 20

Distributions of AP Grades for French Language 2004-2006

AP Grade	Old Form Score Distribution	Equated Score Distribution							
		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
1	.2628	.2828	.2828	.2887	.2887	.2828	.2828	.2828	.2828
2	.2147	.2331	.2397	.2271	.2338	.2331	.2397	.2331	.2331
3	.2863	.2808	.2848	.2808	.2793	.2808	.2848	.3018	.3018
4	.1591	.1410	.1330	.1410	.1384	.1410	.1330	.1374	.1353
5	.0772	.0623	.0598	.0623	.0598	.0623	.0598	.0449	.0470

Table 21

Distributions of AP Grades for French Language 2005-2007

AP Grade	Old Form Score Distribution	Equated Score Distribution							
		Unsmoothed		Presmoothed		Postsmoothed		IRT	
		FE	CE	FE	CE	FE	CE	True	Observed
1	.2365	.2459	.2459	.2459	.2459	.2459	.2459	.2510	.2510
2	.2248	.2385	.2455	.2385	.2385	.2385	.2455	.2334	.2334
3	.2960	.2960	.2889	.2960	.3017	.2960	.2889	.3017	.3017
4	.1610	.1455	.1495	.1455	.1397	.1455	.1455	.1473	.1438
5	.0816	.0742	.0701	.0742	.0742	.0742	.0742	.0666	.0701

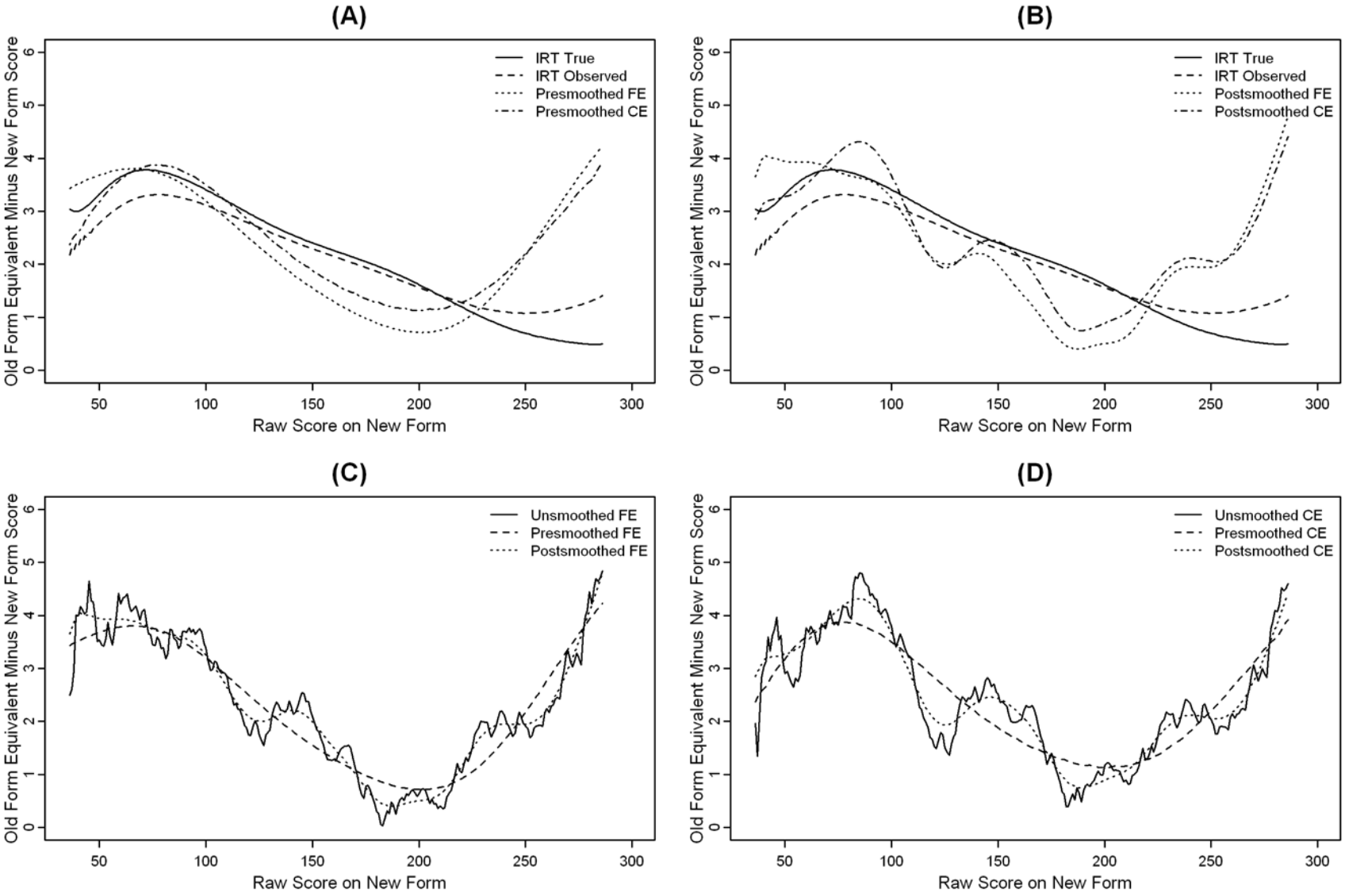

Figure 1. Biology 2004-2006 equating relationships for various equating methods.

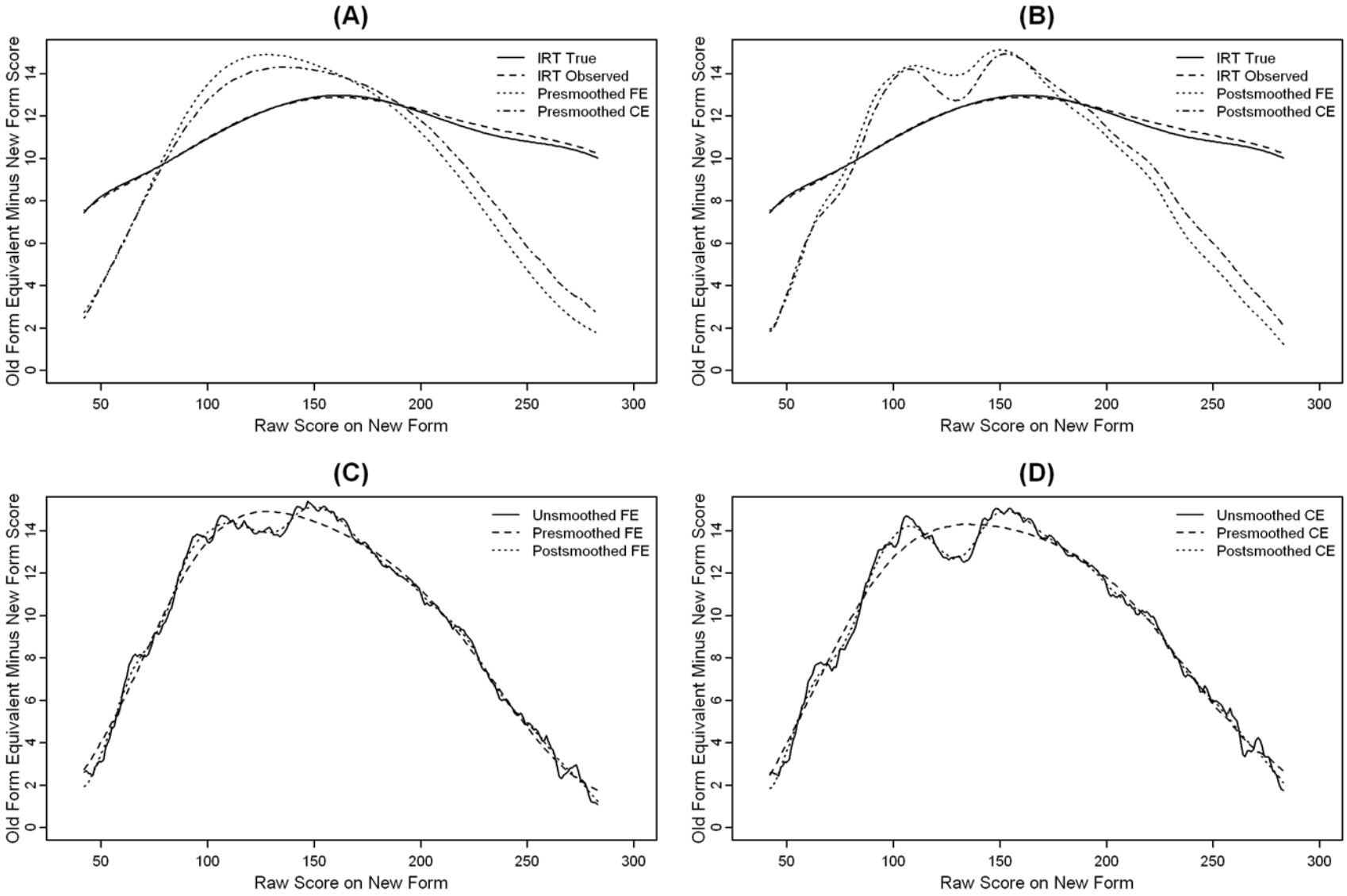

Figure 2. Biology 2005-2007 equating relationships for various equating methods.

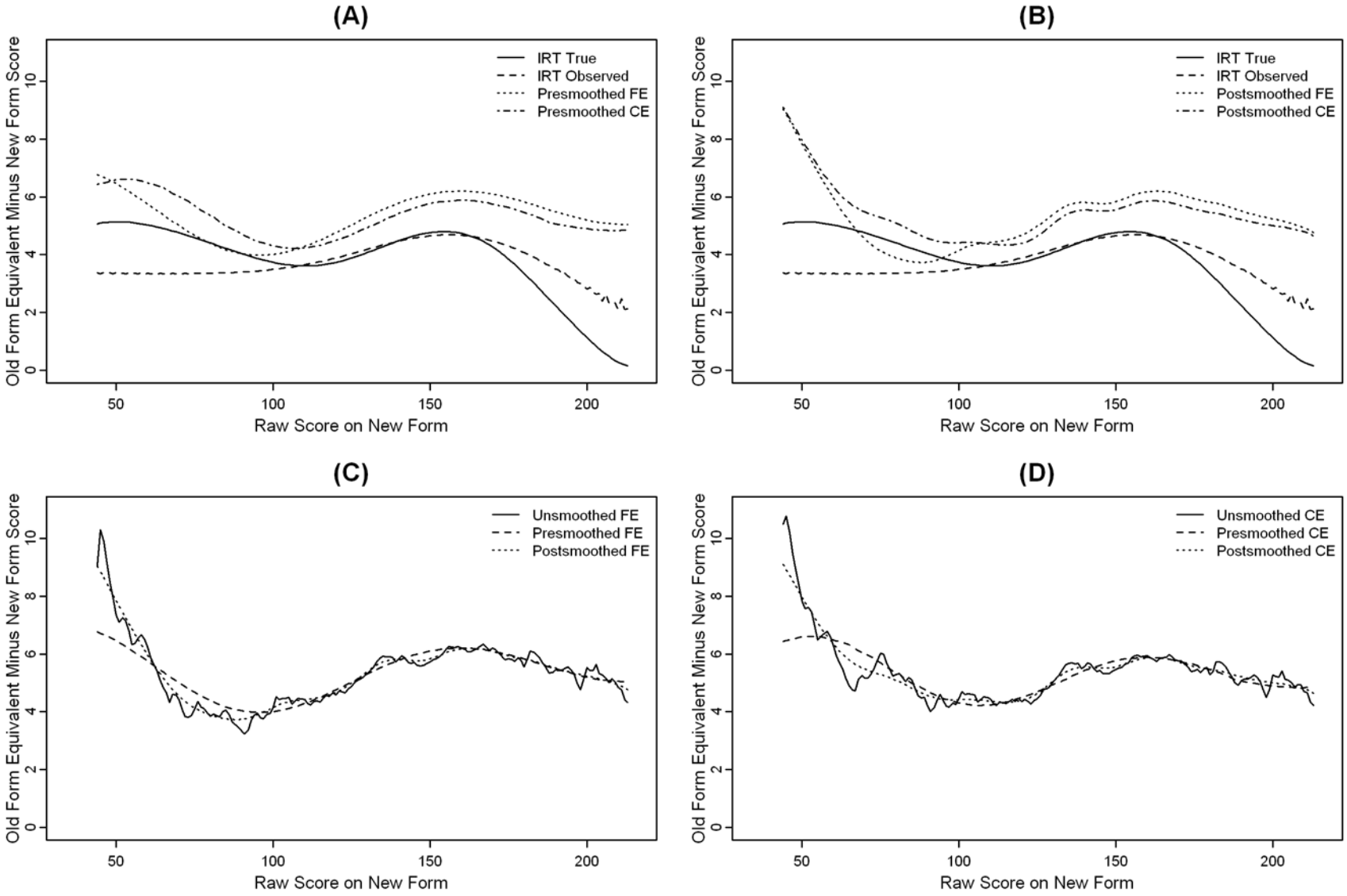

Figure 3. English Language and Composition 2004-2007 equating relationships for various equating methods.

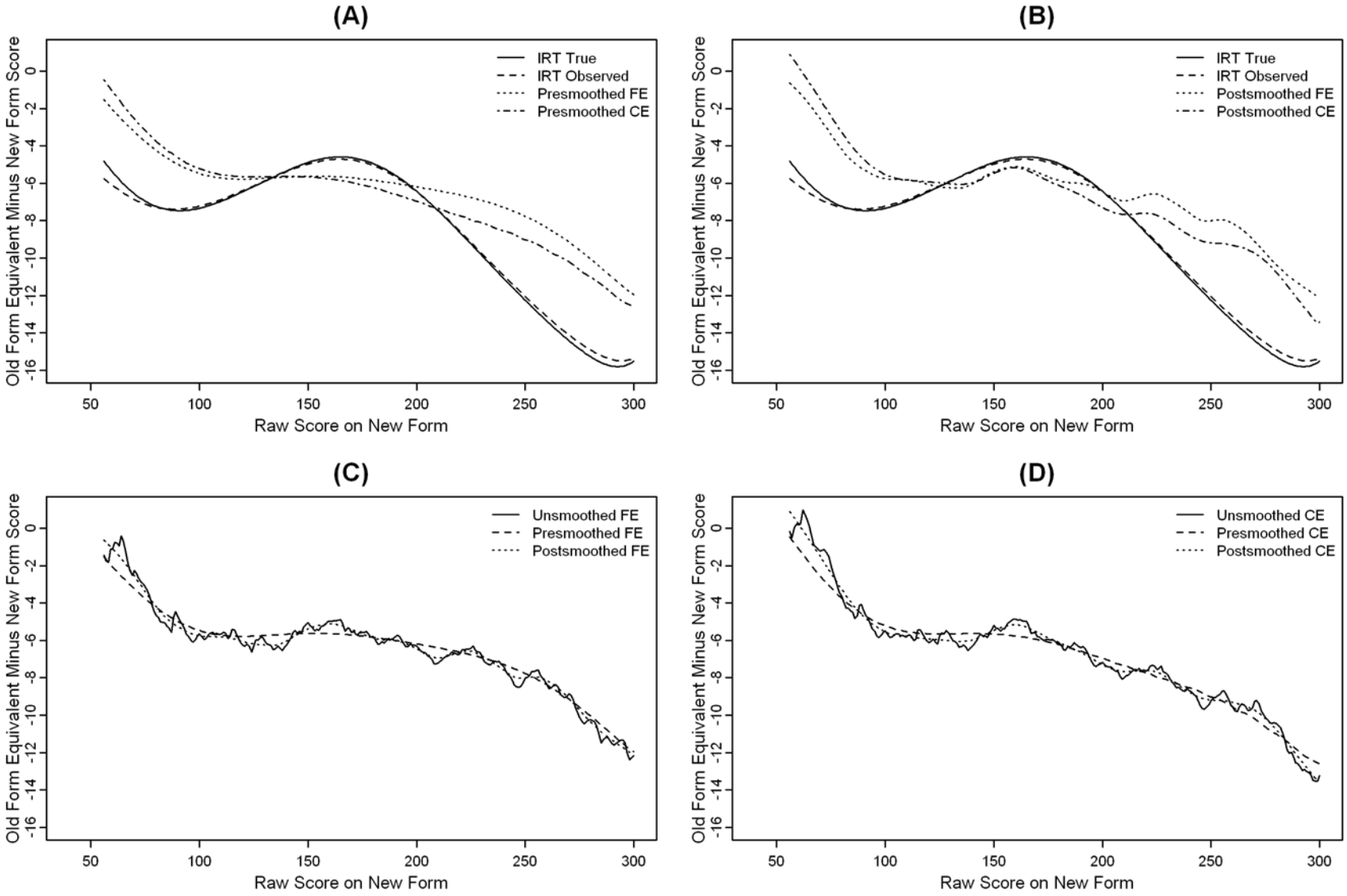

Figure 4. French Language 2004-2006 equating relationships for various equating methods.

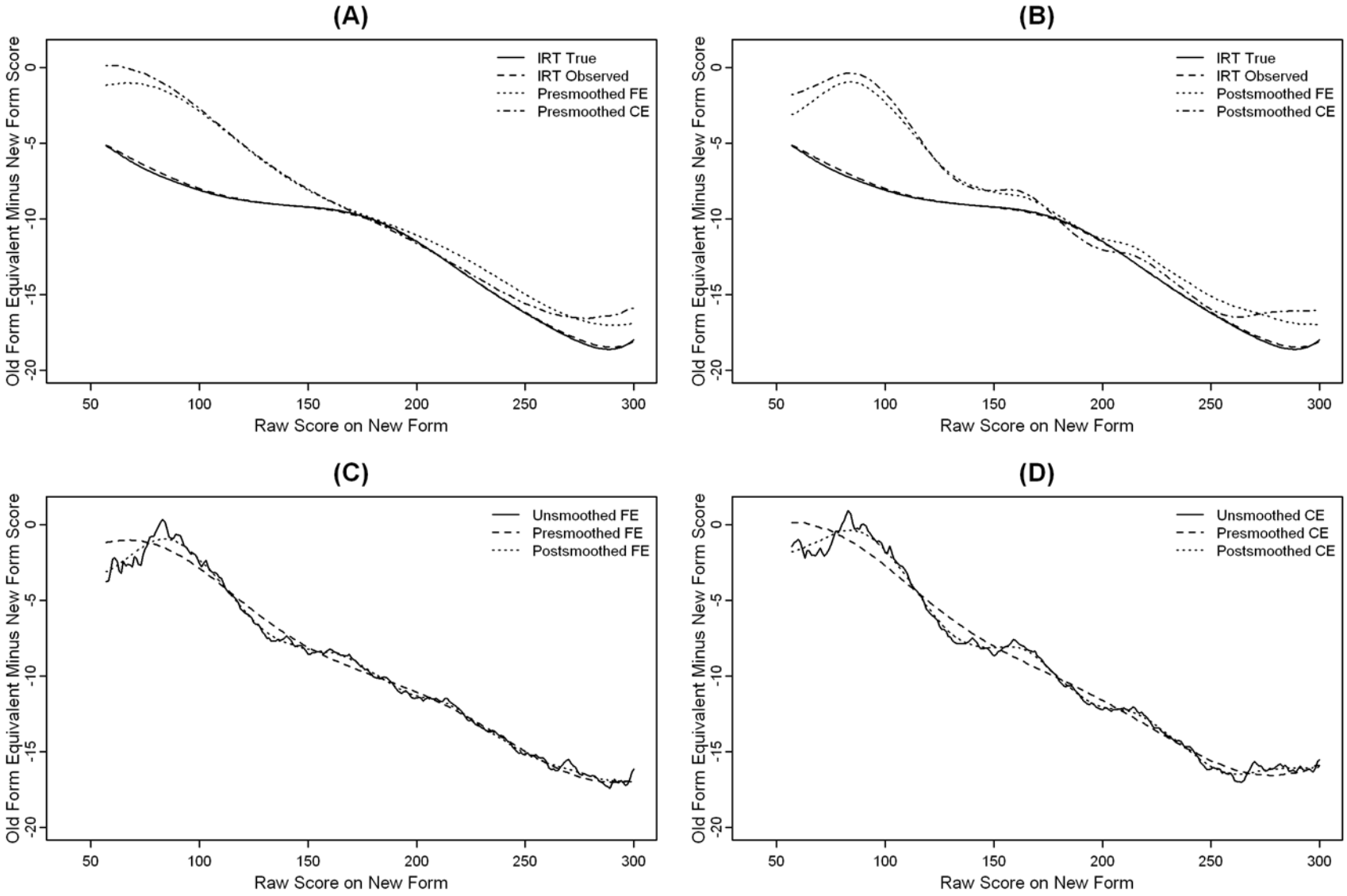

Figure 5. French Language 2005-2007 equating relationships for various equating methods.

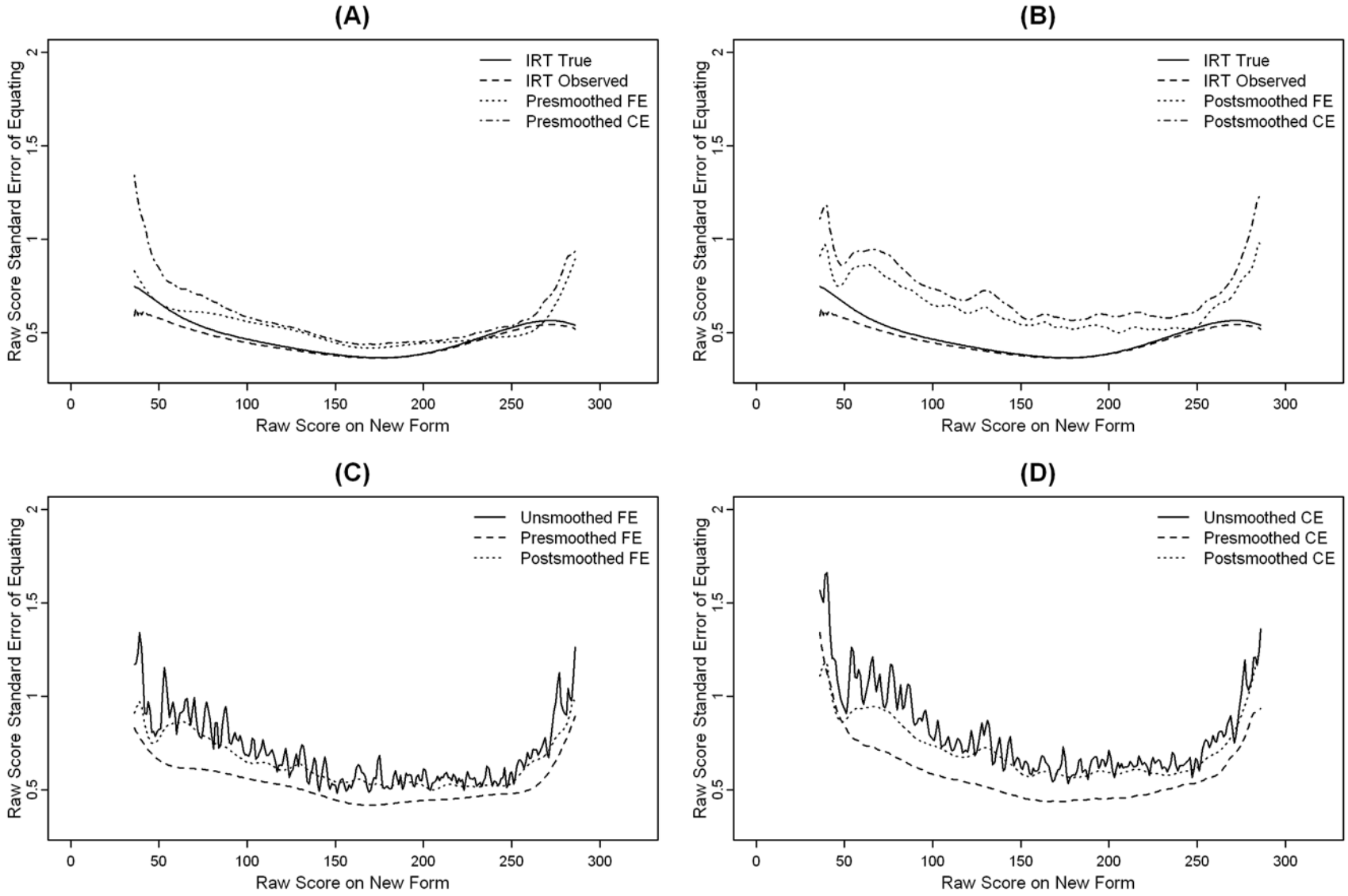

Figure 6. Biology 2004-2006 raw score standard error of equating.

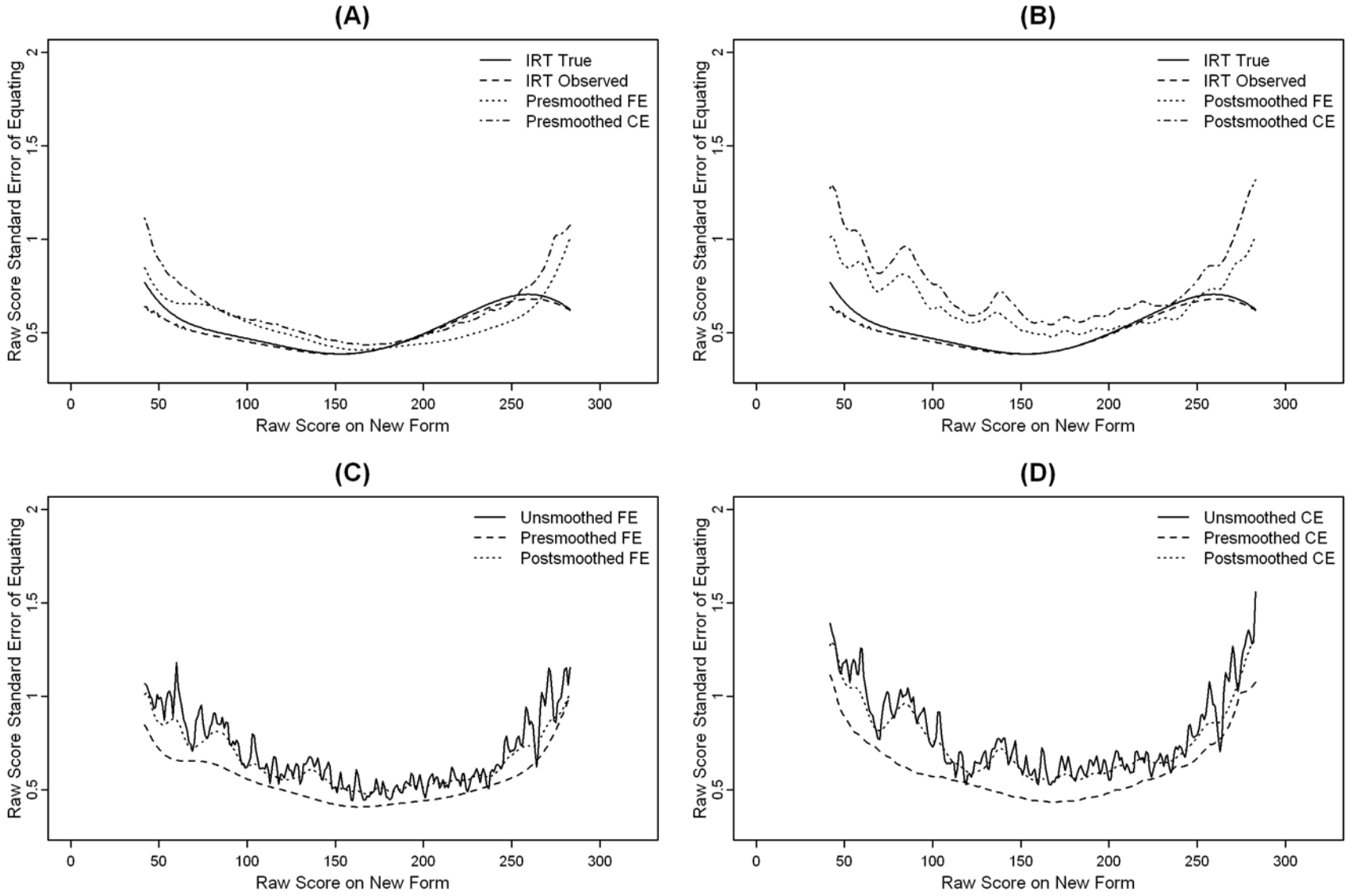

Figure 7. Biology 2005-2007 raw score standard error of equating.

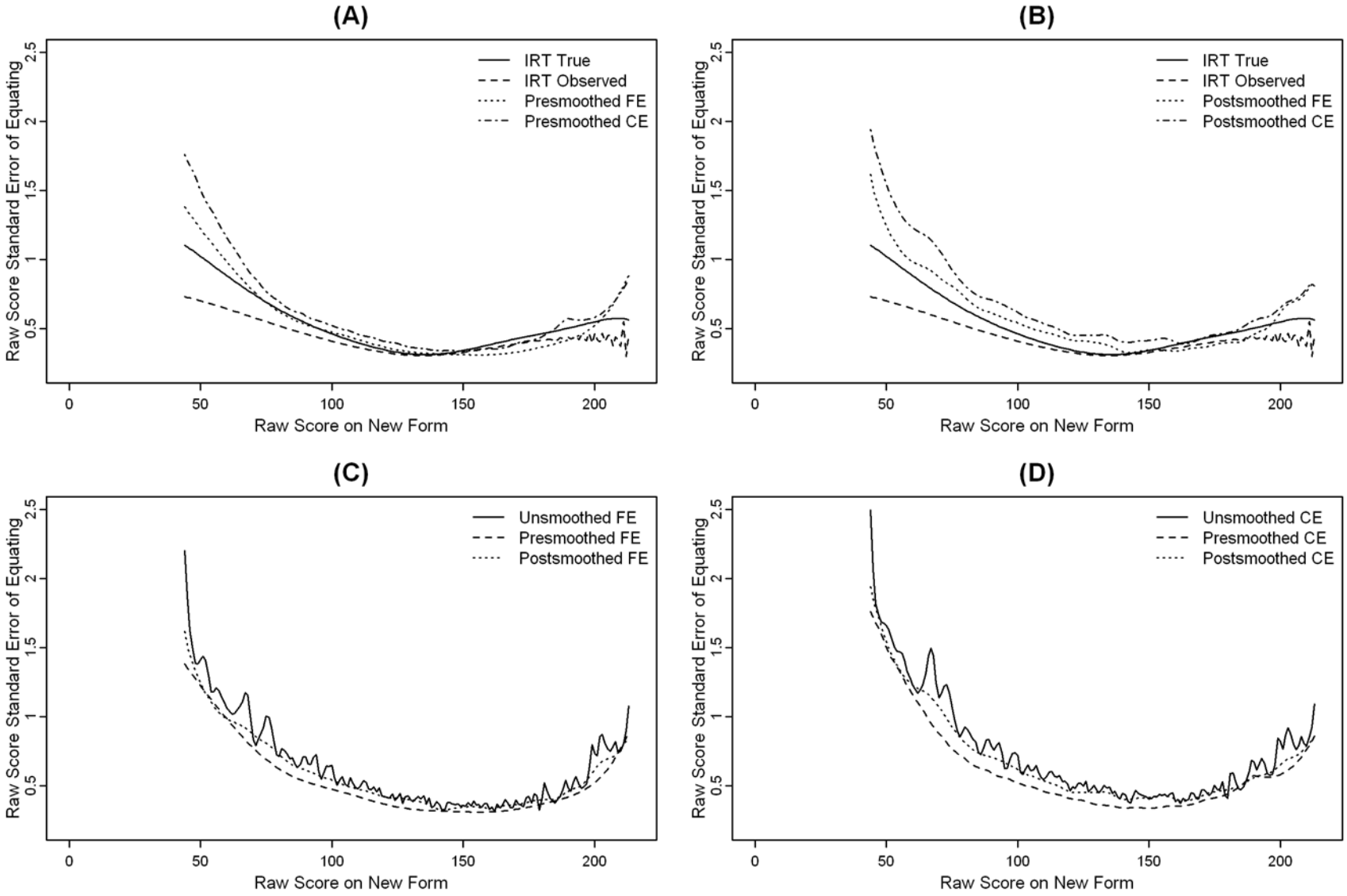

Figure 8. English Language and Composition 2004-2007 raw score standard error of equating.

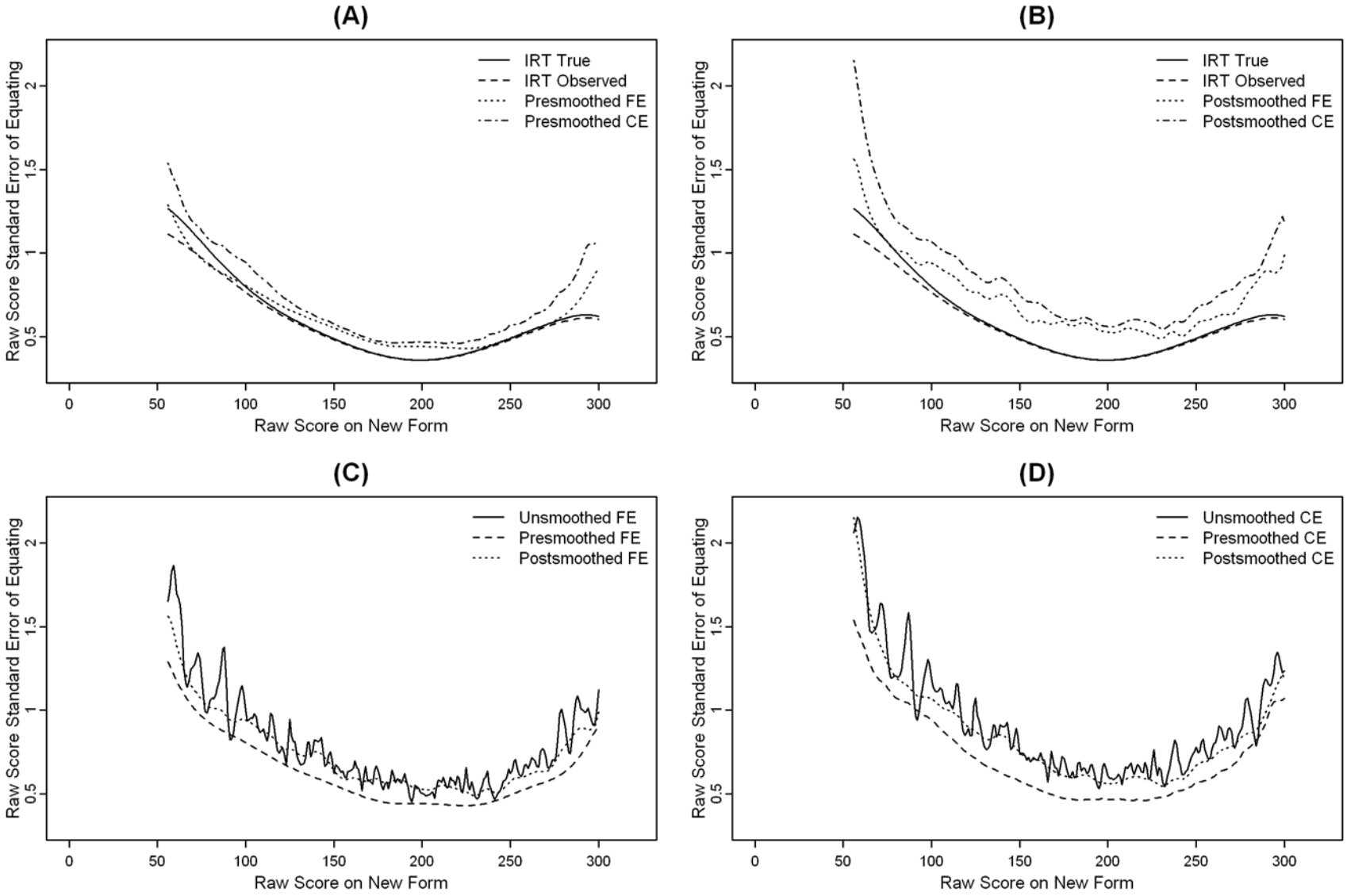

Figure 9. French Language 2004-2006 raw score standard error of equating.

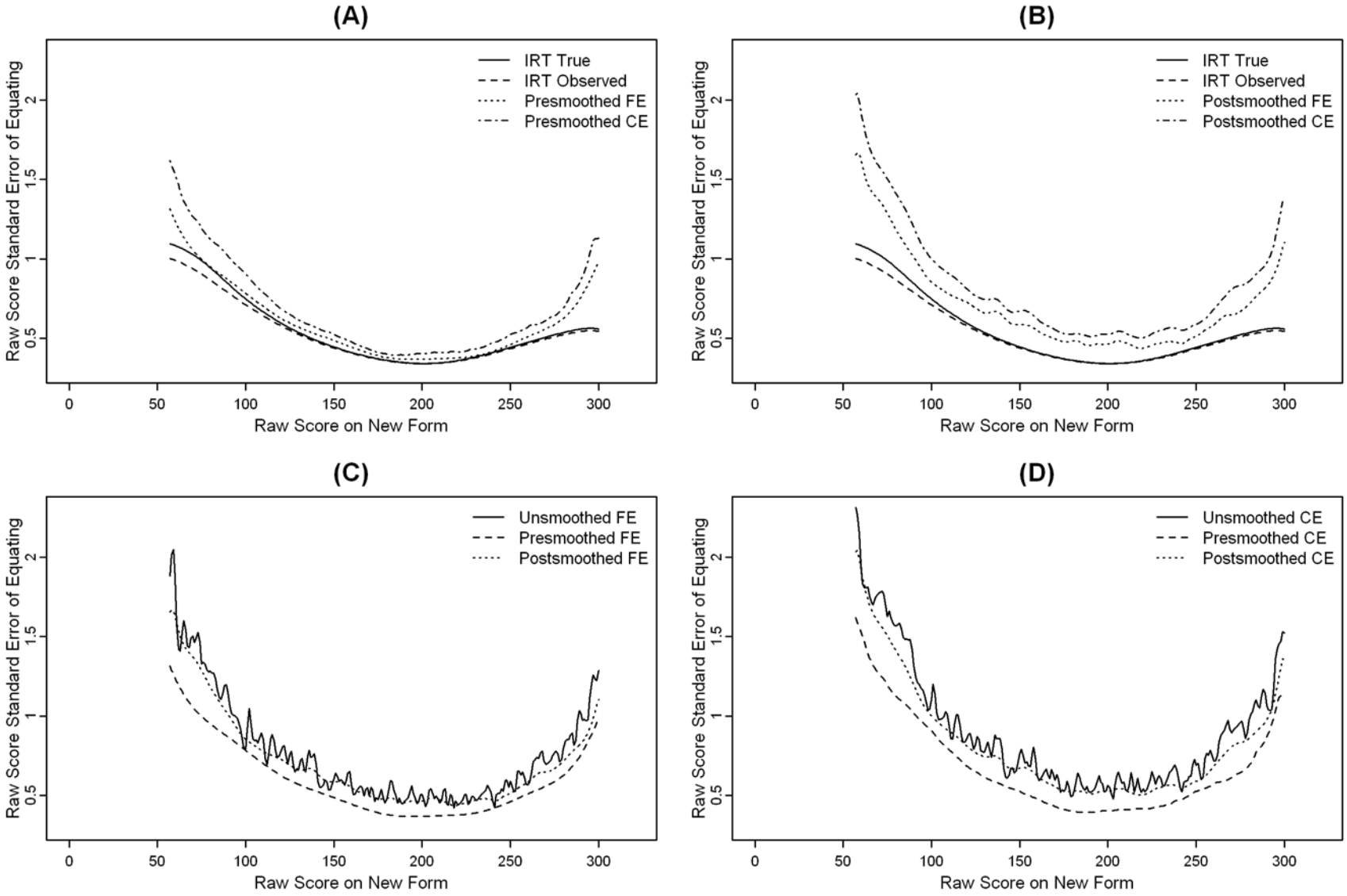

Figure 10. French Language 2005-2007 raw score standard error of equating.

Chapter 3: Effects of Group Differences on Mixed-Format Equating

Sonya J. Powers, Sarah L. Hagge, Wei Wang, Yi He, Chunyan Liu, and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

The purpose of this chapter was to investigate the sensitivity of equating methods to group differences. Data from five Advanced Placement (AP) Exams were used, and for each AP exam, old and new form groups with five common item effect sizes (0, .05, .15, .25, and .5), characterizing five levels of group differences, were created by sampling examinees based on a reduced fee indicator. Equatings were conducted using four methods: frequency estimation (FE), chained equipercentile (CE), item response theory (IRT) true score equating, and IRT observed score equating. In addition, presmoothing and postsmoothing procedures were used for the FE and CE methods. The results showed that as group differences increased, systematic equating error increased for the FE method, whereas the CE method or the IRT methods were not sensitive to group differences. Additionally, for all effect sizes, the IRT methods produced less systematic equating errors than the traditional methods.

Effects of Group Differences on Mixed-Format Equating

In Chapter 2, four equating methods, frequency estimation equipercentile (FE), chained equipercentile (CE), IRT true score, and IRT observed score methods, were compared for mixed-format tests. Although the equating relationships produced by the IRT methods were somewhat different from those produced by the equipercentile methods, the two IRT methods tended to produce results that were similar to each other, and the two traditional methods tended to produce results that were similar to each other. One possible explanation of the observed similarity in equating relationships across different methods is that the groups taking the old and new forms of each exam differed little in ability across the annual administrations for the three AP Exams used in the previous study. However, group differences might exist for other exams or on other occasions. In this chapter, the sensitivity of equating methods to group differences is investigated.

According to Kolen (1990), under the common item nonequivalent groups (CINEG) design, all equating methods tend to produce similar equating functions when there are small group differences and when the common-item set represents the entire test in terms of content and statistical properties. von Davier, Holland, and Thayer (2004) showed mathematically that the FE and the CE methods lead to equivalent results when the groups taking the old and new forms are equivalent or when the common-item scores and total test scores are perfectly correlated. This finding helps to explain why, in Chapter 2, the equating relationships produced by the two traditional equipercentile methods were so similar to each other. However, when group differences are large, equating results produced by different methods might be noticeably different from one another. Kolen and Brennan (2004) suggested that the CE method might be preferable to the FE method when the groups differ substantially because the CE method does not have an explicit requirement that the new and old form groups be very similar. This suggestion was supported by results from Marco, Petersen, and Stewart (1983) and Livingston, Dorans, and Wright (1990).

In a series of studies comparing various equating methods under different matched sampling designs, the sensitivity of equating methods to group differences was found to differ from study to study. Between the FE and CE methods, Cook, Eignor, and Schmitt (1990) found that the FE method was more affected by group differences than the CE method, whereas studies by Lawrence and Dorans (1990) and Schmitt, Cook, Dorans, and Eignor (1990) found the

opposite. Comparing the traditional methods with the IRT methods, the study by Schmitt et al. (1990) showed that the IRT true score method was more affected by sample variation than either FE or CE, whereas Eignor, Stocking, and Cook (1990) found that CE was more sensitive to group difference than the IRT true score method. In addition, a study by Livingston et al. (1990) showed that the sensitivity of equating methods to group differences differs across exams. These studies offer little resolution on which equating method is more robust to group differences. Wright and Dorans (1993) contended that the accuracy of equating results for different methods is affected by the type of sampling design (i.e., representative sampling, new-form matched sampling, reference or target matched sampling).

Using equating a test to itself as a criterion, von Davier, Holland, and Thayer (2003) found that the CE method was slightly less sensitive to group differences than the FE method. Two studies (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008) using IRT-based simulation and/or pseudo-test data showed that FE was more biased than CE when large group differences were present.

In summary, there is some evidence that when groups differ considerably, the CE method may be preferable to the FE method because the CE method is less sensitive to group differences. Most of these studies used single-format tests. Not many studies using mixed-format tests have been performed. The study in this chapter aims to investigate how equating results are affected when mixed-format tests are equated and old and new form groups vary in ability, using our curvilinear equating methods.

Method

Data and Procedure

The current study used the same AP Exams as in Chapter 2. Specifically, the data for Biology 2004-2006, Biology 2005-2007, English Language 2004-2007, French Language 2004-2006, and French Language 2005-2007 were used, and missing responses in these data were imputed using a two-way procedure described by Bernaards and Sijtsma (2000) and Sijtsma and van der Ark (2003). Then, using the imputed data, old and new form groups were sampled to obtain groups with desired differences in common-item performance. These differences were quantified as effect sizes, where the effect size was calculated as the standardized mean difference between groups on the common items. Table 1 provides the sample sizes of the original data, the imputed data, and the data sampled for the current study. Within

one equating link, the same sample size was used for different effect sizes.

Instead of sampling examinees using common-item scores directly, a reduced fee indicator, known to be related to exam performance, was used to sample examinees from the old and new form groups. Five effect sizes (0, .05, .15, .25, and .5), characterizing five levels of differences in common-item scores, were obtained by sampling examinees based on whether or not they received a reduced exam fee. For these sampled examinees, the imputed responses of multiple-choice (MC) items were number-correct scored, and summed scores were used for free-response (FR) items. The same section weights (see Table 2 in Chapter 2) were applied for MC and FR items to form composite scores.

Four equating methods, FE, CE, IRT true score, and IRT observed score, were used to equate scores on the new form (Form X) to scores on the old form (Form Y) for each effect size and each equating link. Altogether 100 equatings (five equating links, five effect sizes, and four equating methods) were conducted. In addition, for the two traditional equipercentile methods, bivariate log-linear presmoothing (Kolen & Brennan, 2004) and cubic spline postsmoothing (Kolen & Brennan, 2004) procedures were applied. For presmoothing, the composite scores were modeled using a polynomial of degree 6, the common item scores with a polynomial of degree 6, and one cross-product was included. The postsmoothing parameter used in the analyses was .1, which was the highest value that resulted in a smoothed equating relationship within error bands of plus and minus one raw score standard error between percentile ranks of .5 and 99.5. For the FE and IRT observed score equating methods, the synthetic weight of .5 was assigned to the group of examinees taking the new form.

For IRT equating, item parameters were estimated using MULTILOG (Thissen, Chen, & Bock, 2003). The three-parameter logistic (3PL) model and the graded response model (GRM) were assumed to estimate item parameters for the MC and FR items respectively. MULTILOG default priors were also used for the item estimation process. Because MULTILOG does not provide an estimate of the posterior ability distribution, which is necessary for IRT observed score equating, quadrature points and weights from a standard normal distribution were used. After obtaining item parameter estimates and quadrature weights, STUIRT (Kim & Kolen, 2004) was used to transform the new form item parameters and ability distribution to the old form scale using the Haebara method (Haebara, 1980) for scale transformation. For all four equating methods, equating and smoothing were implemented using *Equating Recipes* (*ER*; Brennan,

Wang, Kim, & Seol, 2009).

Evaluation Criteria

Equating functions resulting from different effect sizes and different methods were compared graphically. In addition, a modified root expected mean squared difference statistic (REMSD; Dorans & Holland, 2000) was used to quantify differences between the criterion and comparison equating relationships:

$$REMSD = \frac{\sqrt{\sum_{\min(x)}^{\max(x)} v_{xc}[eq_{Yc}(x) - eq_{Y0}(x)]^2}}{\sigma_0(Y)} \quad (1)$$

In this equation, *c* refers to comparison equating relationship, $eq_{Yc}(x)$ is the old form equivalent for the comparison equating relationship, $eq_{Y0}(x)$ is the old form equivalent for the criterion equating relationship with an effect size of 0, v_{xc} is the conditional proportion of examinees at a given *x* score in the comparison group, and $\sigma_0(Y)$ is the old form standard deviation for the criterion group.

Because equating is known to provide reasonable results when group differences are small, the criterion equating relationship was the relationship with an effect size of 0. Each equating method had its own criterion equating relationship and set of comparison equating relationships. Comparison equating relationships included all equatings where the effect size was greater than 0.

Results

Table 2 provides the expected and observed effect sizes and REMSD values that were calculated for each equating method and effect size combination. As effect size increases, it was expected that REMSD might increase, which would indicate sensitivity of the equating method to group differences. This hypothesis was confirmed for all the equating and smoothing methods for French Language 2005-2007 except the unsmoothed CE method. For the other four equating relationships, in general, the results tended to be consistent with this hypothesis only for the FE method as the effect size increased from .15 to .50, but not for the CE or IRT equating methods. The REMSD values with the CE or IRT methods were actually smaller for some of the larger effect sizes. For example, for the two IRT methods used in Biology 2005-2007, the REMSD values associated with an effect size of approximately .50 were smaller than those with lower

effect sizes. For the same exam using the CE method, the REMSD value with effect size of approximately .25 was lower than those with effect sizes of approximately .05 and .15.

A comparison of the average effect size across conditions indicates that the FE method had the highest average REMSD values, followed by CE. IRT average REMSD values were the smallest. Smoothing seems to slightly lower the REMSD values, and presmoothing appears to lower the REMSD values to a slightly greater extent than postsmoothing does. For the two IRT equating methods, the average REMSD values were very similar across the five equating links. IRT observed score equating had slightly smaller average REMSD values than IRT true score equating, except for the relationship of Biology 2004-2006.

Figures 1 to 5 provide comparisons of equating results across effect sizes for postsmoothed equipercentile methods and IRT methods for each of the five equating relationships. Because the REMSD values in Table 2 indicate that the type of smoothing made little impact on the results, the patterns depicted in Figures 1 to 5 would be very similar for unsmoothed or presmoothed traditional methods. The two vertical lines in each figure represent approximately the .5 and 99.5 percentiles of the distribution. Beyond the .5 and 99.5 percentiles, the equating relationships are based on linear interpolation. The vertical axis value of zero represents the criterion equating relationship (ES = 0). The colored lines represent the difference between each of the comparison equating relationships (ES > 0) and the criterion equating relationship. The solid black lines that are symmetric about the vertical axis value of zero provide plus and minus one standard error of equating. Standard errors of equating provide an indication of how much variation between equating relationships can be expected based on sampling error alone. Differences between the criterion and comparison equating relationships that exceed the standard error bands may represent differences beyond what would be expected from sampling error. Because bootstrap standard errors of equating were not calculated for the IRT equating methods, the solid black lines are not included for these methods in Figures 1 to 5.

The results provided in Figures 1 to 5 are consistent with the REMSD values. That is, the deviation of the comparison equating relationships from the criterion equating relationship tended to increase with the increase of effect size across equating methods for only French Language 2005-2007. For the other four equating relationships, this pattern tended to appear for the FE method, but was not consistent for the CE or IRT methods. These results suggest that the FE method is more sensitive to group differences than the CE or IRT methods.

Figures 6 to 10 provide comparisons of equating results across equating methods for various effect sizes for each of the five equating relationships. Because the results were consistent across smoothing conditions, only four methods (postsmoothed FE, postsmoothed CE, IRT true, and IRT observed) were plotted. The two vertical lines in each figure represent approximately the .5 and 99.5 percentiles of the distribution. Beyond these lines, the traditional and IRT true score equating relationships are based on linear interpolation and the IRT observed score equating relationships are based on very little data. The colored lines represent the difference between the old form equivalents and the new form scores for each of the four equating methods. If old and new forms were of equal difficulty across the score scale, all the colored lines would fall on the horizontal line at the vertical axis value of 0. As can be seen from Figures 6 to 10, all the new forms and old forms differed to varying extents. The closer the colored lines are to one another, the more similar the equating results were across equating methods. When the effect size was 0, the equating results for the two traditional methods were very similar to each other, and the equating results for the two IRT methods were very similar to each other. However, as the effect size increased, the traditional equating results became less similar to each other, yet the two IRT equating methods still produced similar results. The CE results were closer to the IRT results, which might be because the IRT and CE methods appear less sensitive to group differences than the FE method.

Summary and Discussion

In this study, sensitivity of equating methods to group differences was investigated. Group differences were simulated by selecting students based on a reduced fee indicator for five pairs of AP Exams. The criterion equating was the equating with no group differences. The results showed that as group differences increased, systematic equating error tended to increase for the FE equating method. However, there did not seem to be a consistent relationship between the magnitude of group differences and systematic equating error for the CE method. This finding is consistent with the previous research (von Davier, Holland, and Thayer, 2003; Livingston et al., 1990; Marco et al., 1983; Sinharay and Holland, 2007; Wang et al., 2008). Also, there appeared to be little, if any, relationship between the magnitude of group differences and systematic equating error for the IRT methods. There appeared to be slightly less systematic equating error with the IRT methods than with the CE method. These findings indicate that for mixed-format test equating, when group differences are substantial, the CE or IRT methods are

preferable over the FE method.

This study involved choosing one sample of examinees at each effect size. The use of a single sample might have made it difficult to identify relatively small relationships between effect size and systematic equating error. In future research, many replications could be used which might lead to greater fidelity in assessing the impact of group differences on the accuracy of equating relationships.

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.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Cook, L. L., Eignor, D. R. & Schmitt, A. P. (1990). *Equating achievement tests using samples matched on ability* (College Board Research report No. 90-2). Princeton, NJ: Educational Testing Service.

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2003). Population invariance and chain versus post-stratification equating methods. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program Examinations* (ETS RR-03-27, pp. 19-36). Princeton, NJ: Educational Testing Service.

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

Doran, N.J., & Holland, P.W. (2000). Population invariance and the equitability 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.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3,* 19–36.

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

Marco, G. L., Peterson, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating methods. In D. Weiss (Ed.), *New horizons in testing* (pp. 147-176). New York: Academic.

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.

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.

Sinharay, S., & Holland, P. W. (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.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG (Version 7.03) [Computer software]. Lincolnwood, IL: Scientific Software International.

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.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS RR-93-04). Princeton, NJ*:* Educational Testing Service.

Table 1

Sample Sizes for the Data Used in the Study

Exam	Equating Link	Original Data		Imputed Data		Sampled Data	
		Old Form	New Form	Old Form	New Form	Old Form	New Form
Biology	2004-2006	20,000	20,000	15,075	16,899	2,050	2,050
	2005-2007	20,000	20,000	16,185	16,819	2,050	2,050
English Language	2004-2007	20,000	20,000	15,820	16,882	2,050	2,050
French Language	2004-2006	13,480	15,154	12,349	13,403	1,050	1,050
	2005-2007	14,495	15,287	13,571	13,982	1,050	1,050

Table 2

REMSD across Effect Sizes and Equating Methods

			REMSD							
	Expected	Observed	Unsmoothed		Presmoothed		Postsmoothed		IRT	
	ES	ES	FE	CE	FE	CE	FE	CE	True	Obs
	.00	.013	Criterion Equating Relationship							
	.05	.054	.0532	.0530	.0451	.0410	.0501	.0495	.0383	.0387
Biology	.15	.168	.0431	.0425	.0404	.0285	.0386	.0326	.0190	.0192
04-06	.25	.260	.0751	.0581	.0629	.0424	.0631	.0463	.0238	.0249
	.50	-.501	.0714	.0451	.0630	.0357	.0648	.0397	.0200	.0212
	Average REMSD		.0607	.0497	.0529	.0369	.0542	.0420	.0253	.0260
	.00	-.017	Criterion Equating Relationship							
	.05	.049	.0493	.0647	.0429	.0597	.0441	.0596	.0474	.0446
Biology	.15	.153	.0640	.0550	.0530	.0472	.0603	.0510	.0580	.0552
05-07	.25	.254	.0570	.0419	.0563	.0309	.0530	.0324	.0521	.0480
	.50	.501	.1096	.0570	.1147	.0499	.1066	.0506	.0279	.0258
	Average REMSD		.0699	.0546	.0667	.0469	.0660	.0484	.0464	.0434
	.00	-.002	Criterion Equating Relationship							
	.05	.060	.0726	.0830	.0618	.0691	.0688	.0760	.0546	.0498
English	.15	.137	.0641	.0642	.0486	.0511	.0572	.0502	.0354	.0315
04-07	.25	.247	.0758	.0590	.0661	.0329	.0659	.0437	.0250	.0177
	.50	-.492	.2543	.1534	.2493	.1470	.2530	.1500	.0863	.0801
	Average REMSD		.1167	.0899	.1064	.0751	.1112	.0800	.0503	.0448
	.00	.004	Criterion Equating Relationship							
	.05	.051	.0734	.0749	.0682	.0507	.0670	.0674	.0326	.0314
French	.15	.148	.1467	.1316	.1198	.1093	.1300	.1174	.0591	.0588
04-06	.25	-.245	.0786	.0654	.0675	.0464	.0711	.0572	.0145	.0142
	.50	.514	.1984	.1584	.1688	.1201	.1845	.1419	.0451	.0435
	Average REMSD		.1243	.1076	.1061	.0816	.1132	.0959	.0378	.0370
	.00	-.007	Criterion Equating Relationship							
	.05	-.050	.0542	.0961	.0395	.0653	.0472	.0815	.0193	.0171
French	.15	.157	.0765	.0939	.0713	.0729	.0723	.0739	.0299	.0290
05-07	.25	.243	.1062	.0870	.0956	.0754	.1020	.0751	.0310	.0295
	.50	.490	.1935	.1272	.1740	.1184	.1917	.1149	.0654	.0644
	Average REMSD		.1076	.1010	.0951	.0830	.1033	.0863	.0364	.0350

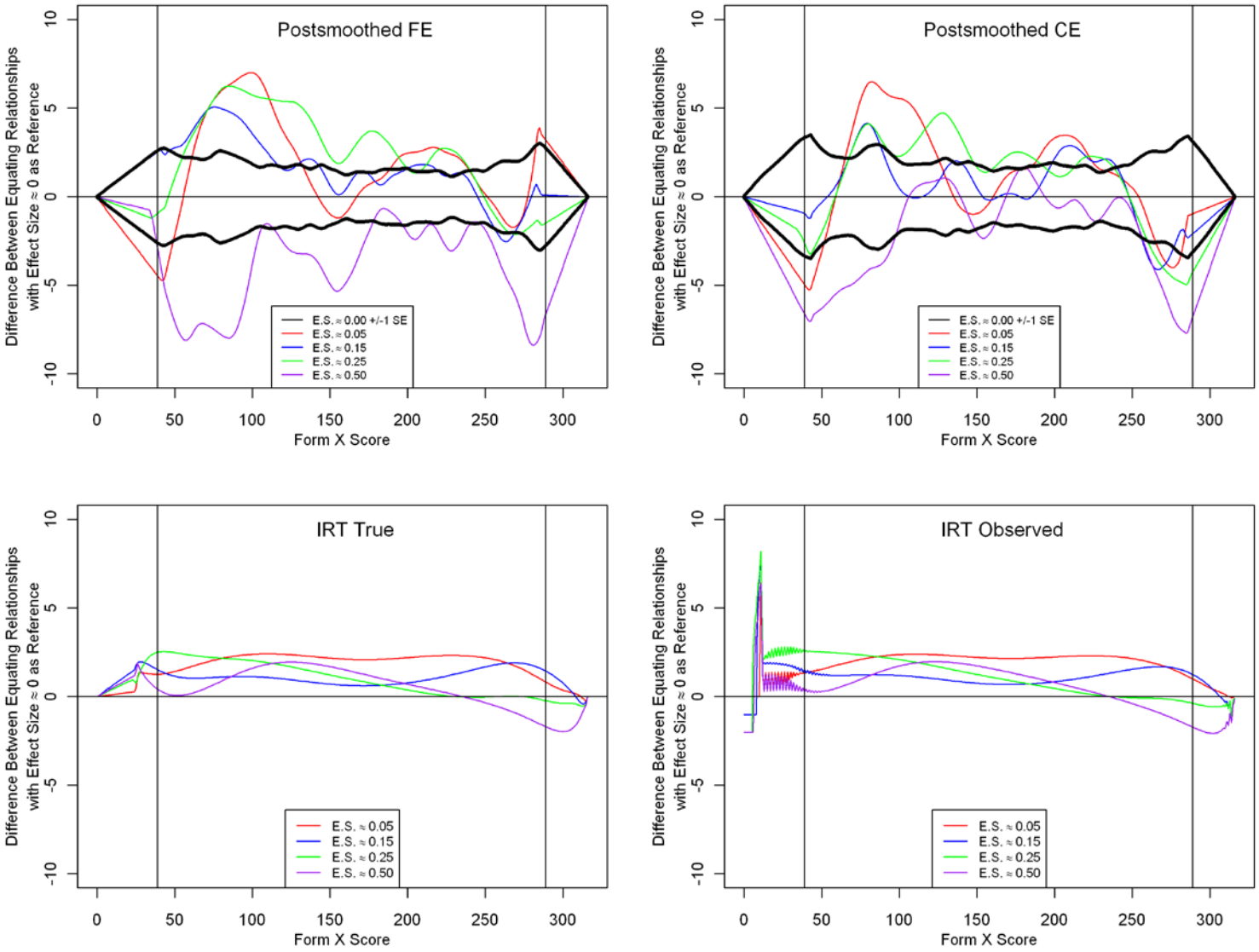

Figure 1. Comparison of equating relationships with varying group differences for Biology 04-06 (Form X: 2006; Form Y: 2004).

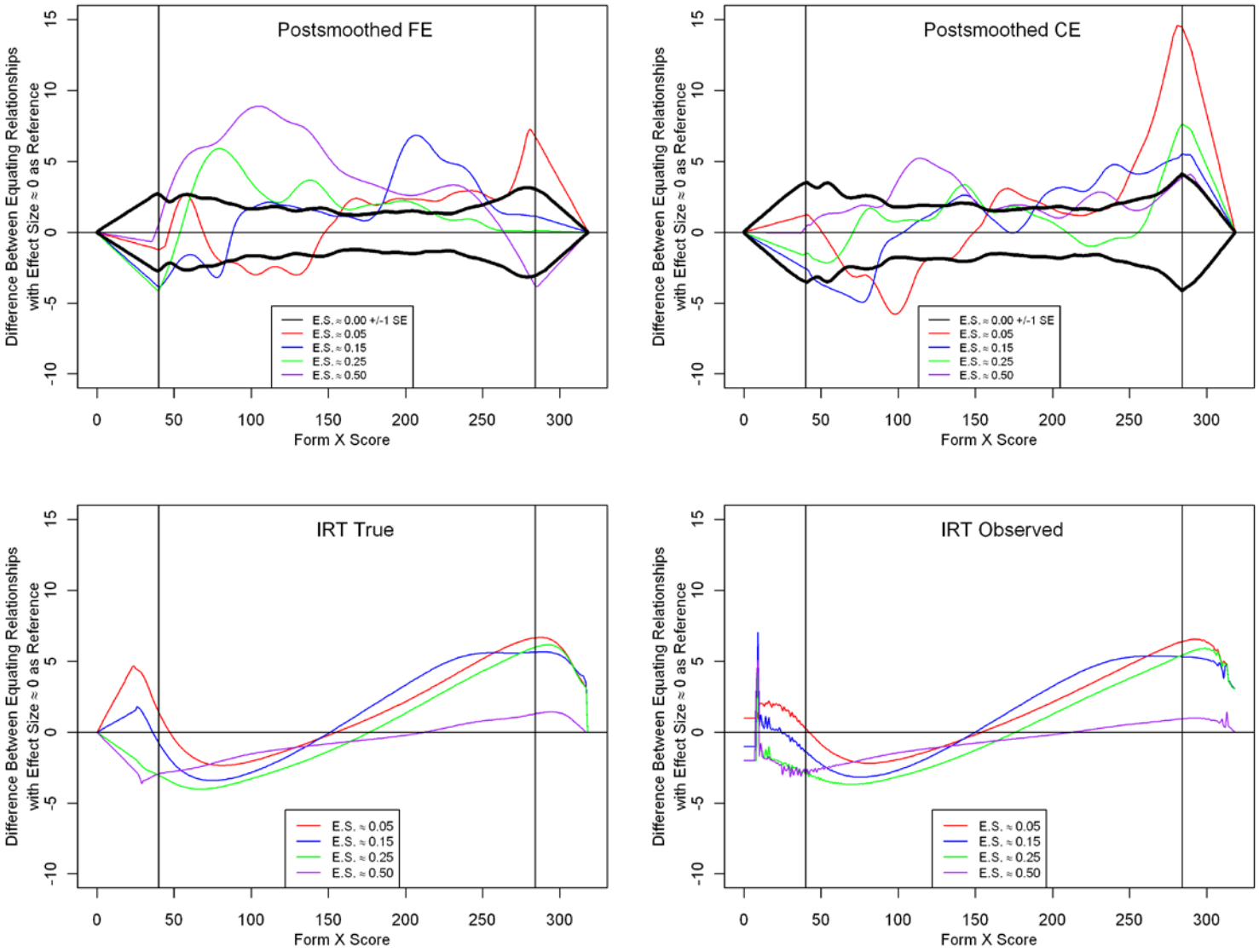

Figure 2. Comparison of equating relationships with varying group differences for Biology 05-07 (Form X: 2007; Form Y: 2005).

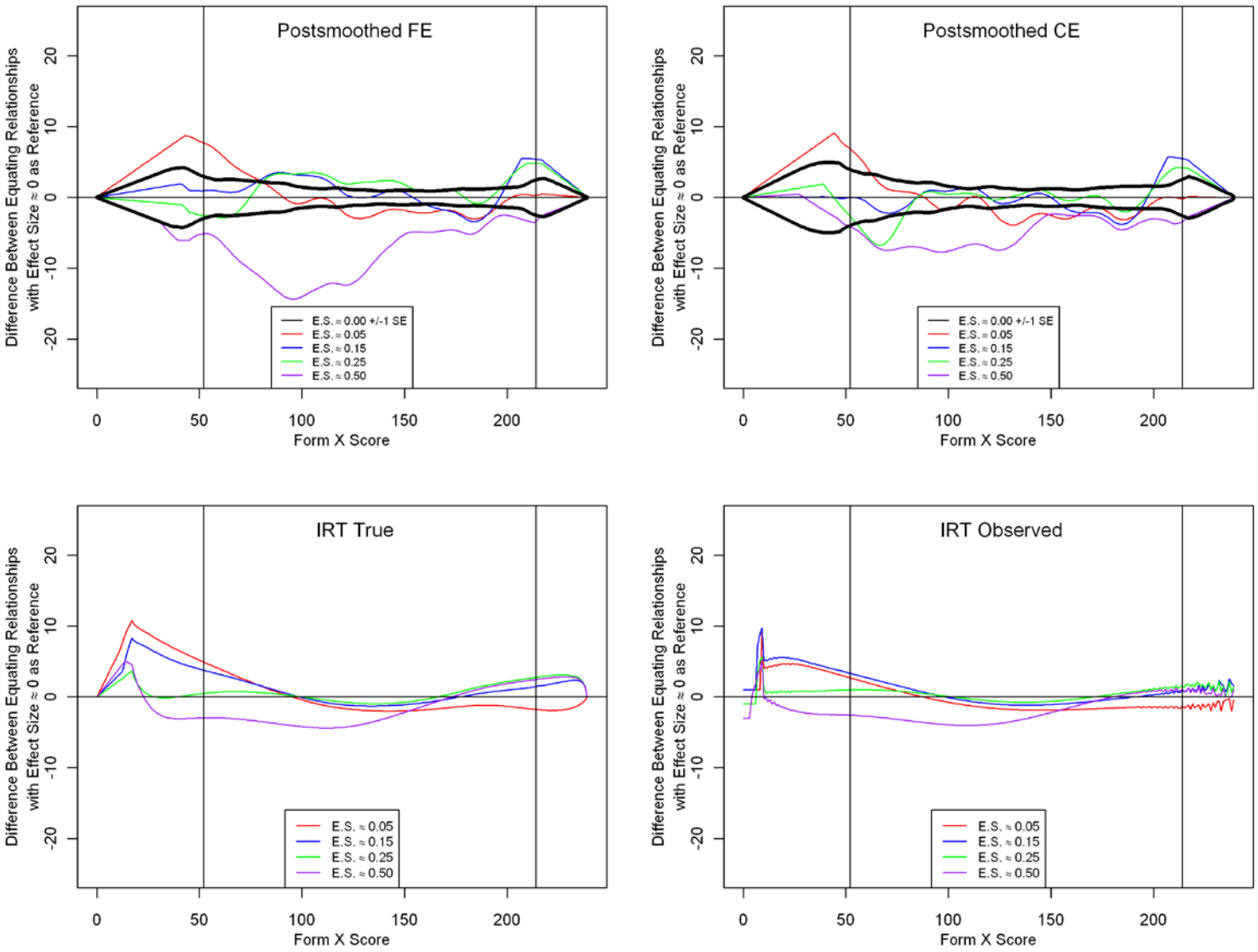

Figure 3. Comparison of equating relationships with varying group differences for English Language 04-07 (Form X: 2007; Form Y: 2004).

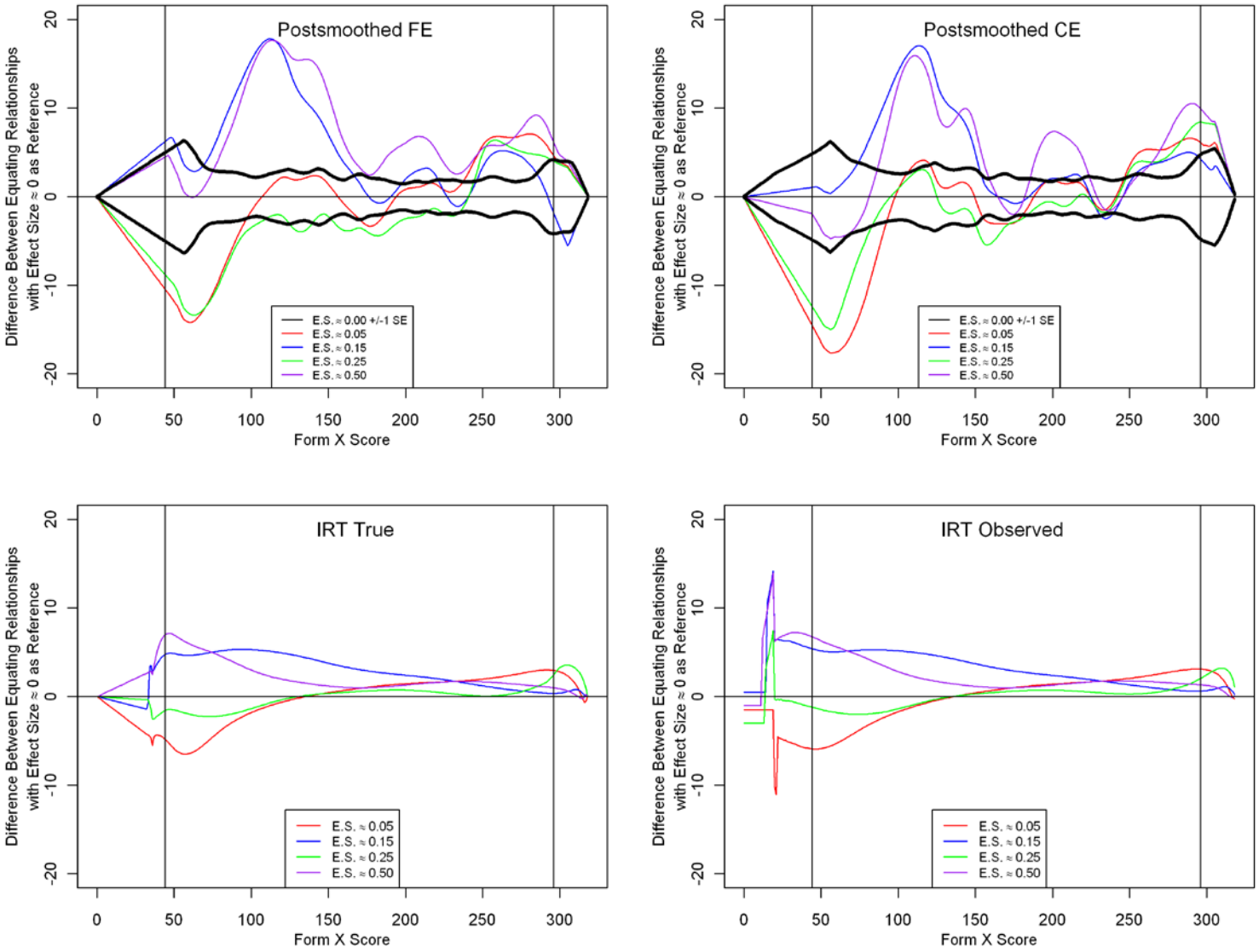

Figure 4. Comparison of equating relationships with varying group differences for French Language 04-06 (Form X: 2006; Form Y: 2004).

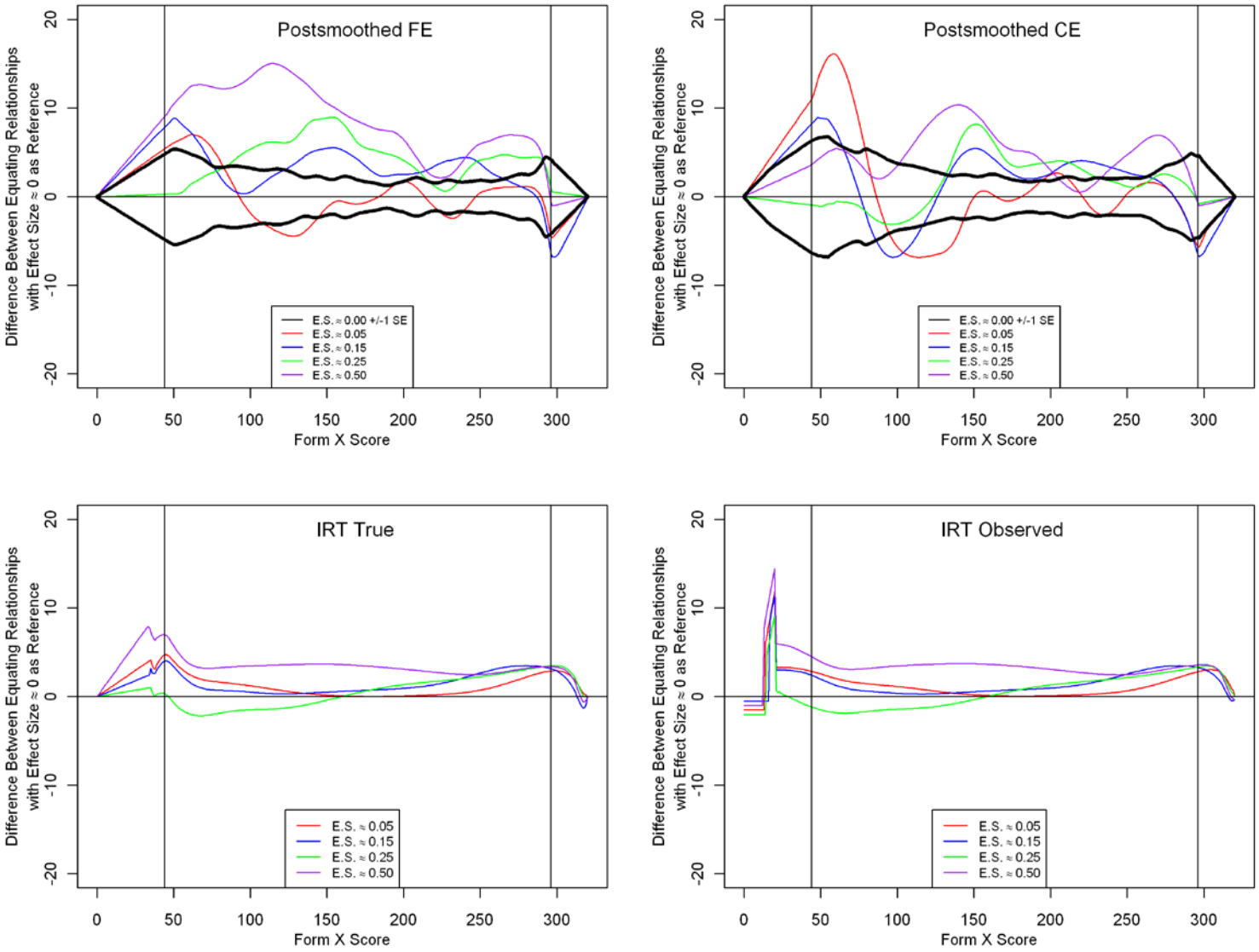

Figure 5. Comparison of equating relationships with varying group differences for French Language 05-07 (Form X: 2007; Form Y: 2005).

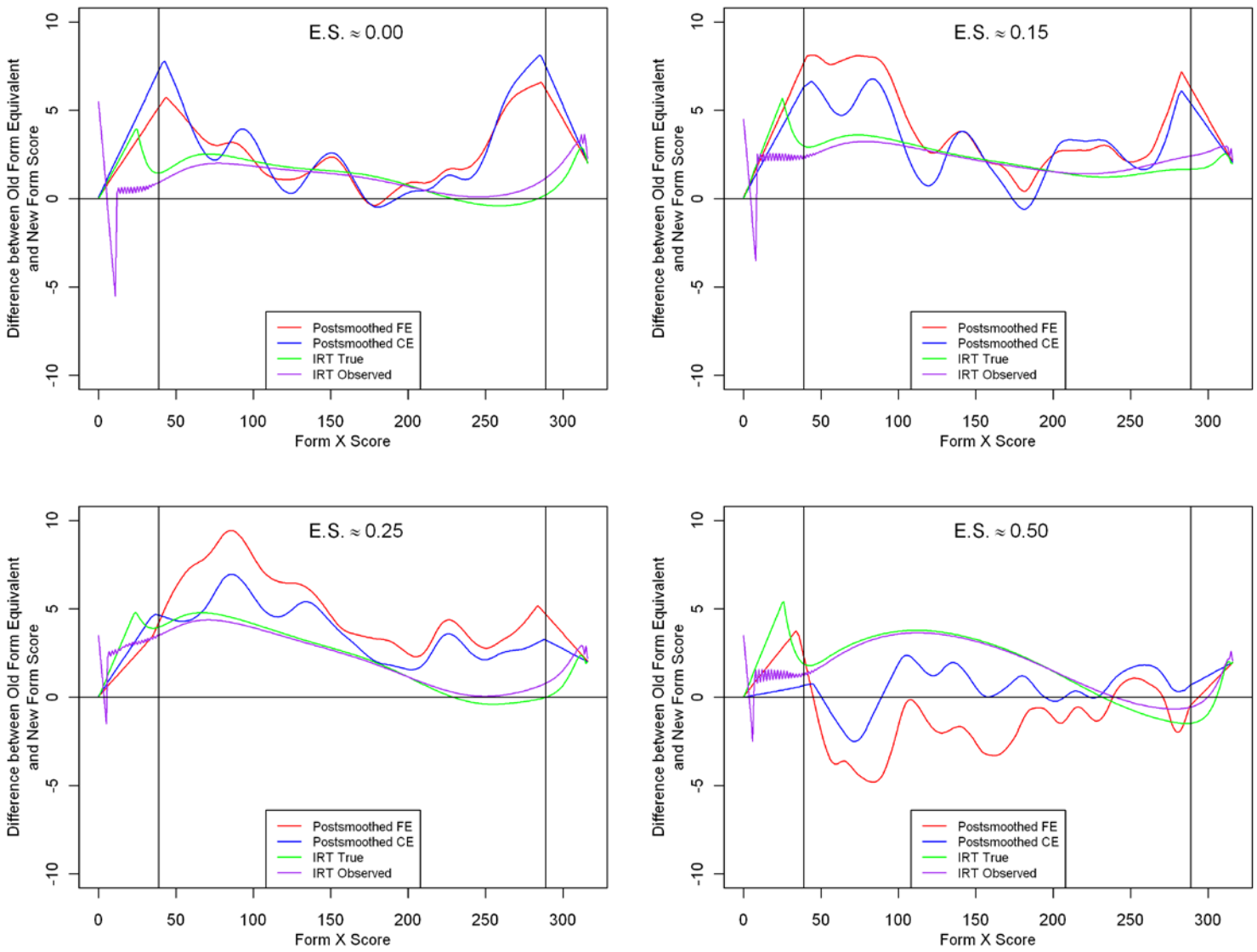

Figure 6. Comparison of Biology 04-06 equating relationships for various equating methods (Form X: 2006; Form Y: 2004).

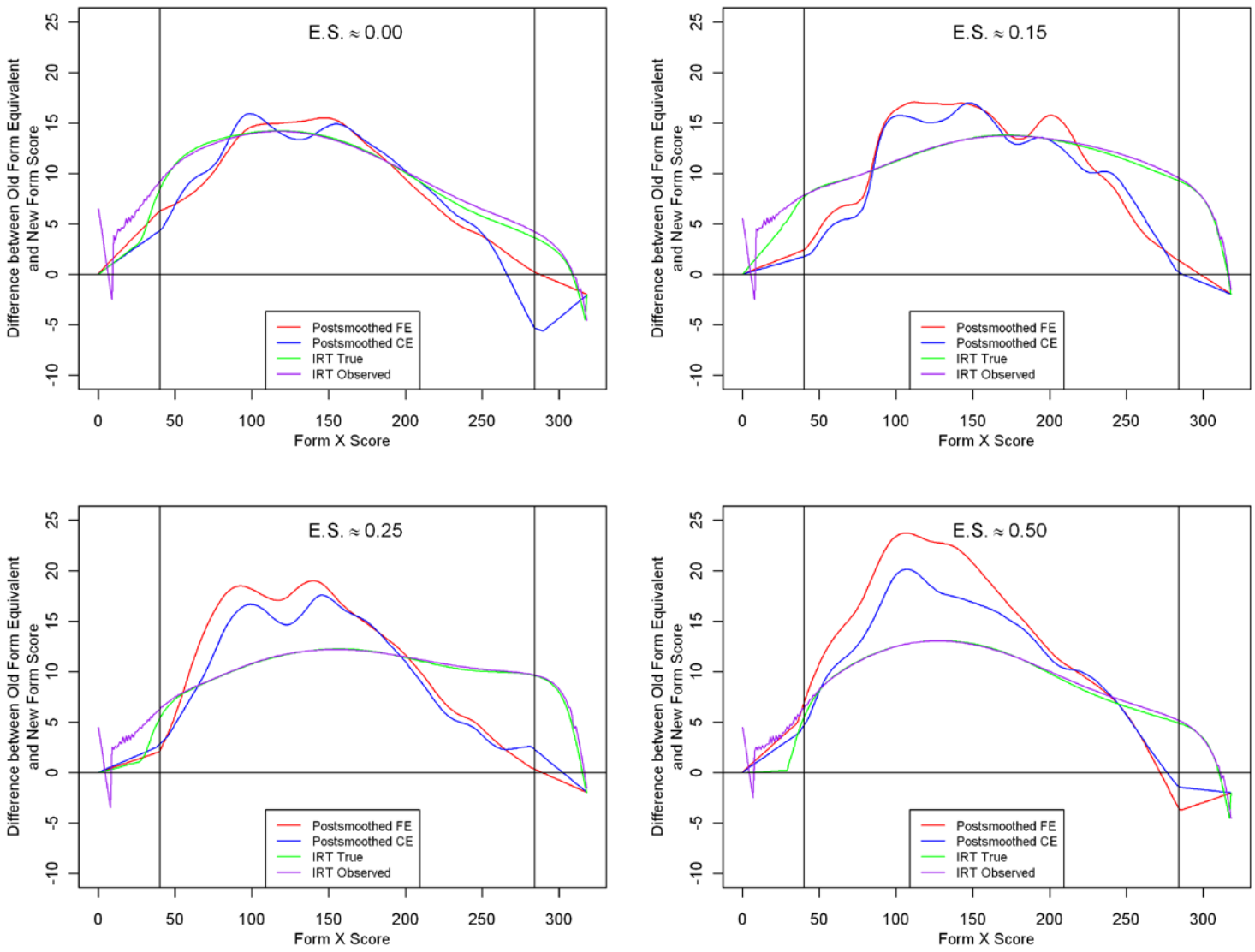

Figure 7. Comparison of Biology 05-07 equating relationships for various equating methods (Form X: 2007; Form Y: 2005).

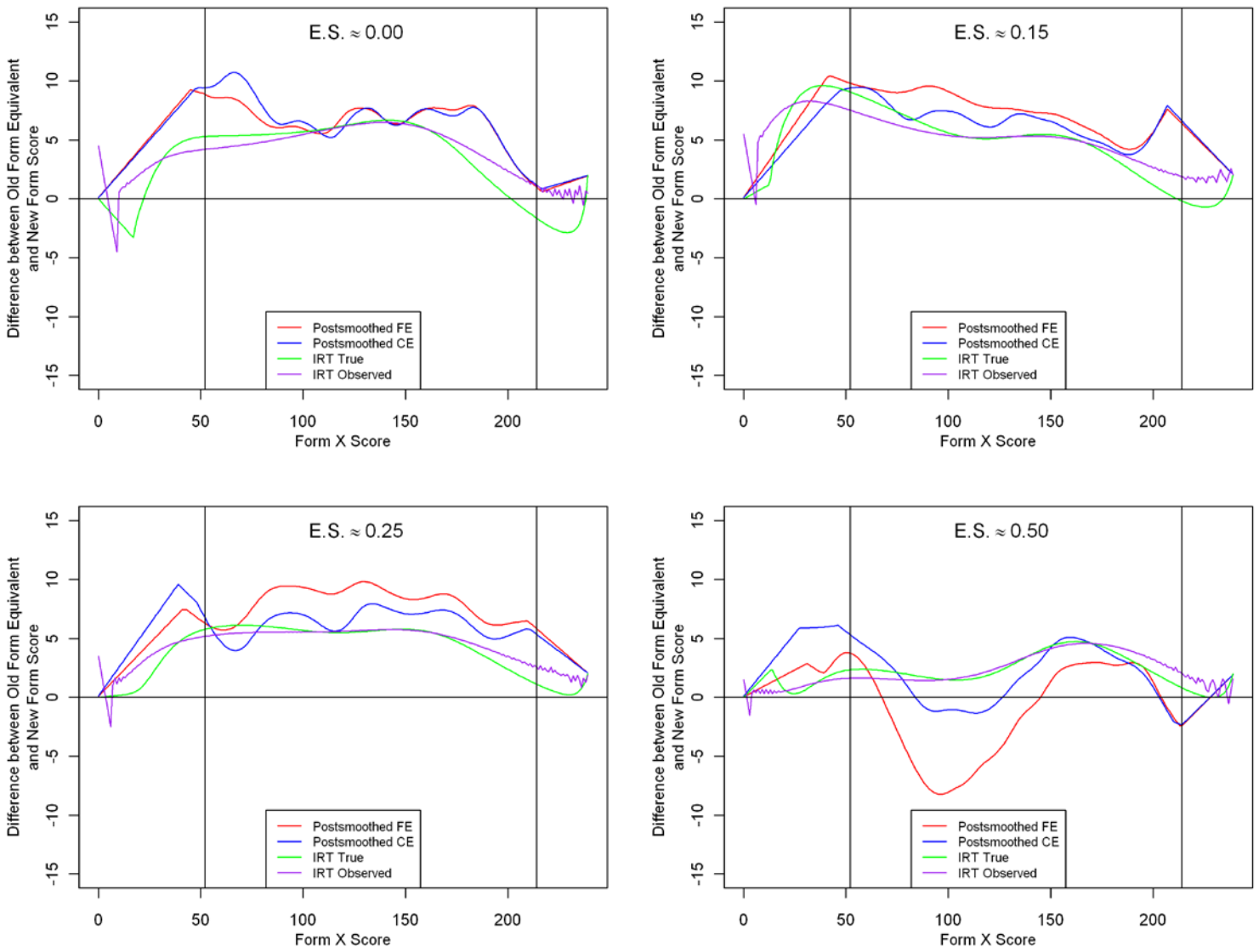

Figure 8. Comparison of English Language 04-07 equating relationships for various equating methods (Form X: 2007; Form Y: 2004).

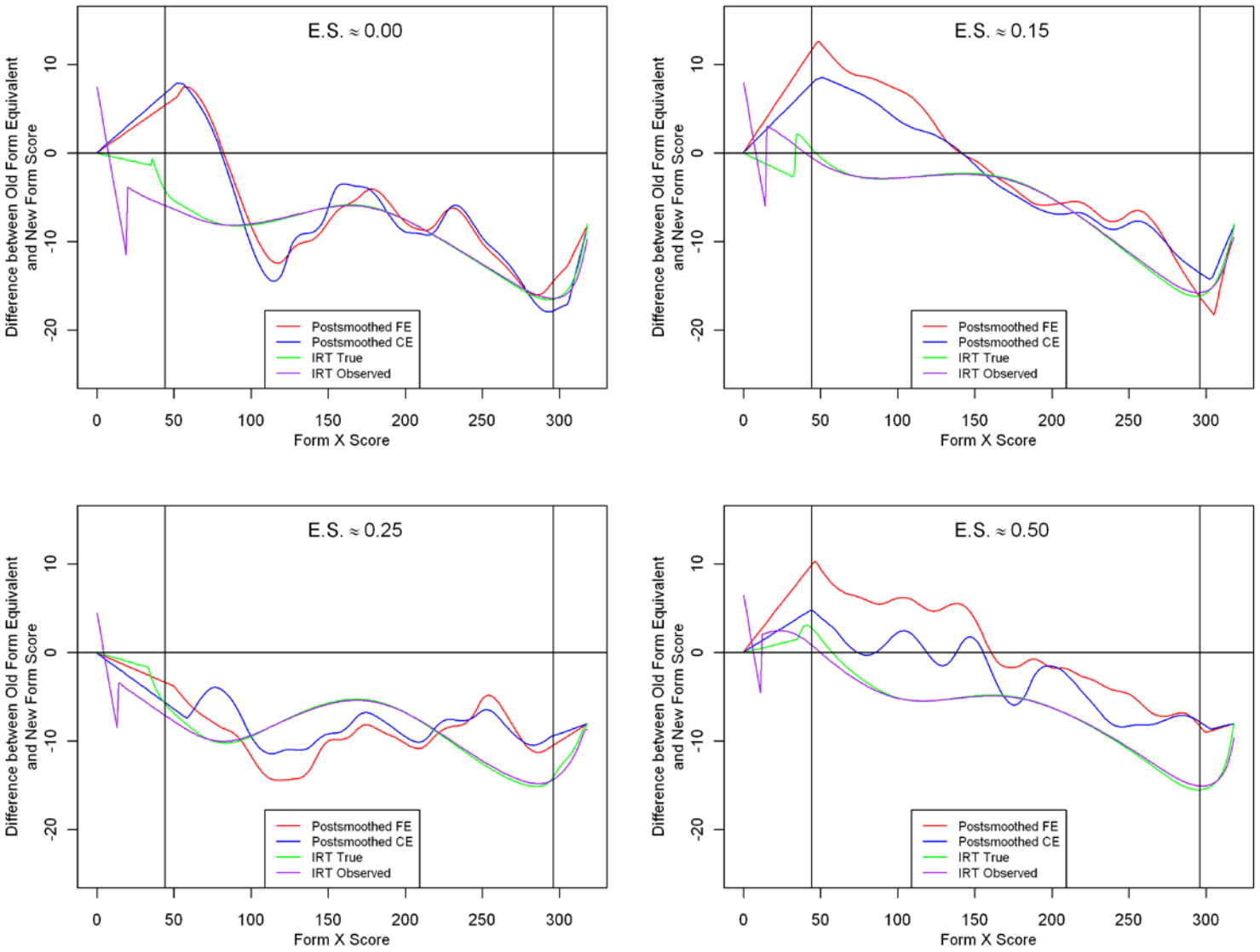

Figure 9. Comparison of French Language 04-06 equating relationships for various equating methods (Form X: 2006; Form Y: 2004).

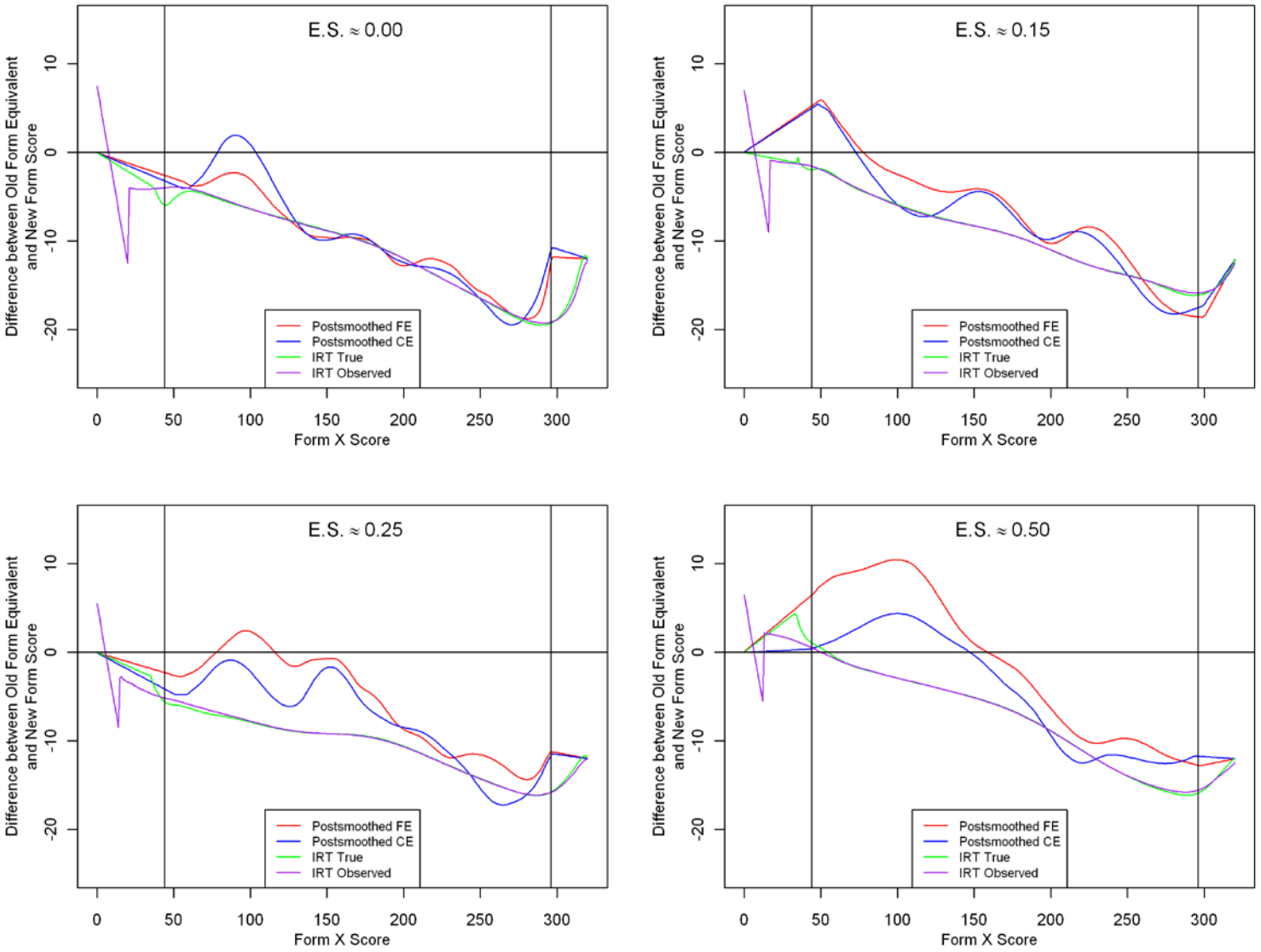

Figure 10. Comparison of French Language 05-07 equating relationships for various equating methods (Form X: 2007; Form Y: 2005).

Chapter 4: A Comparison Among IRT Equating Methods and Traditional Equating Methods for Mixed-Format Tests

Chunyan Liu and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

Many large-scale tests are composed of items of various formats such as multiple choice (MC) and free-response (FR) formats. The use of mixed-format tests complicates equating, especially when using the common-item nonequivalent group (CINEG) design. The results for IRT true and observed score equating methods and traditional frequency estimation and chained equipercentile equating were compared under the CINEG design using two empirical data sets. A pseudo-test design in combination with the bootstrap method was used to compute standard errors of equating, bias, and root mean square errors indices of equating. The equating relationships for the four equating methods were found to be quite different for one test and very similar for the other test. The results for the standard errors of equating were consistent across the two tests. The standard errors of equating for the IRT observed score equating method were slightly smaller than those for the IRT true score equating method which had smaller standard errors of equating than the frequency estimation equipercentile method. The chained equipercentile method had the largest standard errors of equating. The results for the bias and root mean square error of equating indices suggested that the IRT observed score equating method had smaller biases and root mean square error indices than the IRT true score equating method. Other than that finding, the results for the bias and root mean square error indices were not consistent across the two tests.

A Comparison Among IRT Equating Methods and Traditional Equating Methods for Mixed-Format Tests

Many large-scale tests are composed of items of various formats. For example, some items on a test might be multiple choice (MC) and others free response (FR). The FR items are typically scored by raters, and they generally have different numbers of score points from the MC items. Also, because of security concerns, the FR items are sometimes not used as common items. All these factors complicate the equating of mixed-format tests, especially for the common-item nonequivalent group (CINEG) design where two test forms are administered to two samples from two different populations and the common items in the two forms are used to link the scores. One guideline for assembling the common item set is that the common item set should be proportional to the total test forms in content and statistical characteristics (Kolen & Brennan, 2004). However, the common item set for mixed-format tests often contains only MC items.

Item response theory (IRT) has been used in many applications. One of the main applications of IRT is equating including IRT true score equating and IRT observed score equating. For mixed-format tests, the raw score on a test is often found by summing the item scores for the different item types, which are referred to as summed scores. Under IRT true score equating, the true summed score on the new form is equated to the true summed score on the old form. One disadvantage of IRT true score equating is that the true scores are not available in practice. Under IRT observed score equating, the estimated item parameters and proficiency distributions are used to estimate the distributions of observed summed scores for new and old forms. Then traditional equipercentile equating is used to equate the new form to the old form. The distribution of proficiency in the population of examinees is required for IRT observed score equating. When using IRT to conduct equating, it is often suggested that traditional equating methods also be conducted as a check if possible (Kolen & Brennan, 2004).

IRT equating methods and traditional equating methods have been compared in various research studies (e.g., Cook, Dunbar, & Eignor, 1981; Han, Kolen, & Pohlmann, 1997; Harris, 1991; Kolen, 1981; Petersen, Cook, & Stocking, 1983). Many studies have been conducted to estimate the standard errors of equating for the traditional methods (Cui & Kolen, 2009; Kolen, 1985; Liou, Cheng, & Johnson, 1997; Wang, Lee, Brennan, & Kolen, 2008; Zeng, Hanson, & Kolen, 1994). However, due to the computational intensity, only a few studies (Cho, 2007; Liu,

Schulz, & Yu, 2008; Tsai, Hanson, & Kolen, 2001) examined bootstrap standard errors of equating (*SEE*) using IRT equating methods. Little research has been conducted to compare the *SEE* of IRT equating methods and traditional equating methods for mixed-format tests.

Pseudo-test forms have been used to address research questions that cannot be addressed by the operational test forms because of the lack of evaluation criterion (von Davier et al., 2006; Holland, Sinharay, von Davier, & Han, 2008; Kim, Walker, & McHale, 2008, 2010; Ricker & von Davier, 2007; Sinharay & Holland, 2007, 2008, 2010; Walker & Kim, 2009). Pseudo-test forms are created by splitting one operational test form into two similar test forms. Because all examinees take all of the items in the original operational form, they have scores on both pseudo-test forms. Therefore, a single group data collection design can be used for the pseudo-test and a criterion equating relationship can be created. In addition, the researcher can easily manipulate the composition of items on pseudo forms based on the research questions to be addressed.

The objective of this paper is to compare two IRT equating methods (true and observed score equating) and two traditional equating methods (frequency estimation equipercentile equating (FE) and chained equipercentile equating (CE) using the CINEG design for mixed-format tests. More specifically, the following research questions were addressed in this paper:

1. How different are the equating relationships among the four equating methods?
2. Which method produces the smallest random error and bias?

Data

Data for this study were from the College Board Advanced Placement Program® (AP®) tests. It is important to note that, although Advanced Placement (AP) tests were used for analyses in this study, the test data 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 dissertation should not be made to the operational AP tests.

Data from two AP tests were used in this study: AP World History and AP Biology, which contain two types of items, MC and FR items. Formula scoring was used for operational scoring of the MC items for these test forms. In the directions for the examinations, examinees were instructed that they could guess on the MC items but would be penalized for incorrect guessing. For example, .25 points were subtracted from examinee scores if their response to each MC item was wrong during operational scoring. If they did not answer the item, there was no

penalty. Consequently, there were a large number of missing responses on MC items for some examinees in the operational scoring.

In this study, number correct scoring was used for the MC items, which means that correct responses were scored as 1 and incorrect responses were scored as 0. There was no penalty for incorrect responses. Examinees with more than 20% omitted responses were removed from the data sets. For the remaining examinees, a two-way imputation procedure (Bernaards & Sijtsma, 2000; Sijtsma & van der Ark, 2003) was used to impute the scores for MC items with omitted responses.

The World History operational test form contains 70 MC items and 3 FR items. The FR items are scored 0 to 9. Only one form was used in this study and the data set was based on 16,965 examinees tested in 2005. The Form was then split into two pseudo Forms (Form X and Form Y) by even-odd items. Both pseudo forms had 42 MC and 2 FR items (the second FR item appeared in both forms). Every third MC item was a common item and no FR items served as common items (in total there were 14 common items), referred to as V. The composite score for the pseudo forms was the sum of MC item score and twice the FR item score. So the score range is from 0 to 78. Since both pseudo forms (all items) were administered to all 16,965 examinees, the 16,965 examinees were treated as the population. Figure 1 provides the relative frequency distributions for the population who took the History test and Table 1 provides the mean, standard deviation (sd), skewness, and kurtosis of the distributions. In general, the Form X and Form Y for the History test were very similar.

The other data set is from the AP Biology test that contains 98 MC items and 4 FR items. The FR items are scored 0 to 10. The data set contained records for 16,185 examinees who were tested in 2005. The original test was again split into two pseudo forms (Form X and Form Y) by even-odd items (the last three items were not used). Both pseudo forms had 57 MC items and 2 FR items. Every third MC item was a common item and no FR items served as common items (in total there were 19 common items). The composite score of the pseudo forms is also the sum of MC item score and twice of the FR item score. So the score range is from 0 to 97. The 16,185 examinees were treated as the population. Figure 2 provides the relative frequency distributions for the population who took the Biology test, and Table 1 provides the summary statistics of the distributions. In general, the Form X was easier than the Form Y for the Biology test.

To form non-equivalent groups of examinees on the two pseudo forms for each test, 2,000 males and 1,000 females were randomly drawn with replacement from the population and combined to form Group 1. Similarly, 1,000 males and 2,000 females were randomly drawn with replacement from the same population and combined to form Group 2. Group 1 was assigned to take pseudo Form X and Group 2 was assigned to take pseudo Form Y. These data are referred to the sample data in this paper. Using such a procedure, a CINEG design was created. The descriptive statistics, correlation between the composite score and the common item score, and the effect sizes between the common item score means are provided in Table 2.

Method

In this paper, the 3-parameter logistic (3-PL) model was used to estimate the item parameters for the MC items and the graded response model was used to estimate the item parameters for the FR items using MULTILOG (Thissen, Chen, & Bock, 2003). The Stocking-Lord scale transformation method (Stocking & Lord, 1983) was used to transform the item parameters and theta distribution of Form X scale to the Form Y scale. In this section, the equating relationship for the population is described first followed by the description of equating relationships for the sample data, and the steps used in the simulation study.

Two computer programs were used to conduct IRT equating. STUIRT (Kim & Kolen, 2004) is a computer program written in the C language to conduct the scale transformation for mixed-format test as well as for single format tests. POLYEQUATE (Kolen, 2004) is a computer program written in the *FORTRAN* language to conduct IRT true and observed score equating using dichotomous and polytomous IRT models. STUIRT, POLYEQUATE were both downloaded from the CASMA website (http://www.education.uiowa.edu/casma/computer_programs.htm).

Equating Relationship for the Population

For the History test, since all 16,965 examinees took both pseudo test Form X and Form Y, the 16,965 examinees were treated as the population and the design used for population data collection is a single group design. Log-linear presmoothing with C=6 for both forms were used to smooth the population distribution and then equipercentile equating was used to find the equating relationship for the population and these equivalents were treated as the population values ($e_Y(x_i)$) which were used to compute the bias. The population equating relationship for the Biology test was calculated similarly.

Estimation of the Four Equating Relationships for the Sample Data

1. IRT Equating: The item parameters for pseudo Form X and Form Y were estimated separately using MULTILOG. The theta distributions were assumed to be normal for both groups. The STUIRT computer program was then used to find the slope and intercept to put the item parameters and theta distribution of Form X onto the scale of Form Y using the Stocking-Lord method. Finally, the POLYEQUATE computer program was used to find the equated scores for IRT true and observed score equating.
2. Traditional Equating: The composite score and common item score were computed for the two sample data. Frequency estimation equipercentile equating and chained equipercentile equating were conducted using the *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009) open source C functions. Postsmoothing with smoothing parameter *S*=.1 (Kolen & Brennan, 2004) was used to find smoothed equivalents.

Simulation Study

To compute the *SEE* of the four equating methods, begin with the sample data. The data consist of N_X =3,000 examinees' item responses on pseudo Form X and N_Y =3,000 examinees' item responses on pseudo Form Y. The following steps were adapted from Kolen and Brennan (2004) and used to compute the *SEE*:

1. Draw a random bootstrap sample of item responses with replacement of size N_X from the item responses on pseudo Form X;
2. Draw a random bootstrap sample of item responses with replacement of size N_Y from the item responses on pseudo Form Y;
3. Run MULTILOG separately for the samples obtained in steps 1 and 2 to estimate item parameters. The 3-PL model and the graded response model were used for MC items and FR items, respectively. The proficiency distributions were assumed to be normal for both groups.
4. Run the STUIRT computer program to estimate the Stocking-Lord scale transformation coefficients using the common-item parameters estimated in step 3.
5. Apply the slope and intercept obtained in step 4 to rescale the Form X item parameters and proficiency distribution in step 3 to the Form Y scale.

6. Run the POLYEQUATE computer program to obtain the Form Y equivalents of x_i for IRT true- and observed-score equating using the rescaled item parameters obtained by step 5. Refer to these estimates as $\hat{e}_{IRT\text{-}True,1}(x_i)$ and $\hat{e}_{IRT\text{-}Observed,1}(x_i)$, where IRT-True and IRT-Observed represent the IRT true score equating and IRT observed score equating.
7. Compute Form X score and common item scores for the sample in step 1.
8. Compute Form Y score and common item scores for the sample in step 2.
9. Conduct FE and CE equating with postsmoothing (*S*=.1) using the *Equating Recipes* open source C functions. Refer to these estimates as $\hat{e}_{FE,1}(x_i)$ and $\hat{e}_{CE,1}(x_i)$.
10. Repeat steps 1 to 9 1,000 times to get $\hat{e}_{IRT\text{-}True,1}(x_i)$, $\hat{e}_{IRT\text{-}True,2}(x_i),\ldots,\ \hat{e}_{IRT\text{-}True,1000}(x_i)$; $\hat{e}_{IRT\text{-}Observed,1}(x_i)$, $\hat{e}_{IRT\text{-}Observed,2}(x_i),\ldots,\ \hat{e}_{IRT\text{-}Observed,1000}(x_i)$; $\hat{e}_{FE,1}(x_i)$, $\hat{e}_{FE,2}(x_i),\ldots,\ \hat{e}_{FE,1000}(x_i)$; $\hat{e}_{CE,1}(x_i)$, $\hat{e}_{CE,2}(x_i),\ldots,\ \hat{e}_{CE,1000}(x_i)$. The *SEE* for each method was computed by

$$SEE(x_i)=\sqrt{\frac{\sum_{r=1}^{1000}[\hat{e}_{Y,r}(x_i)-\overline{e}_Y(x_i)]^2}{1000-1}}\ , \tag{1}$$

where

$$\overline{e}_Y(x_i)=\frac{\sum_{r=1}^{1000}\hat{e}_{Y,r}(x_i)}{1000}\ \text{, and} \tag{2}$$

Y represents IRT-True, IRT-Observed, FE, or CE.

Bias and *RMSE* were computed using the following formulas:

$$Bias(x_i)=\left|\overline{e}_Y(x_i)-e_Y(x_i)\right|\text{, and} \tag{3}$$

$$RMSE(x_i)=\sqrt{Bias(x_i)^2+SEE(x_i)^2}\ , \tag{4}$$

where $e_Y(x_i)$ is the equated score for the population.

Considering all score points together, the mean standard error of equating (*MSEE*), average bias (*ABIAS*), and average root mean square error (*ARMSE*) are defined as follows:

$$MSEE=\sqrt{\sum_{i=0}^{K}f(x_i)\{SEE(x_i)\}^2}\ , \tag{5}$$

$$ABIAS=\sqrt{\sum_{i=0}^{K}f(x_i)\{Bias(x_i)\}^2}\ \text{, and} \tag{6}$$

$$ARMSE=\sqrt{\sum_{i=0}^{K}f(x_i)\{RMSE(x_i)\}^2}\ , \tag{7}$$

where *K* is the maximum score on the pseudo test and $f(x_i)$ is the population relative frequency of Form X at score point x_i.

Results

In this section, the comparison of the equating relationships, standard errors of equating, biases, and root mean square errors of the four equating methods are presented first for the History test and then for the Biology test.

History Test

Comparison of the equating relationships for the sample data. The equating relationships for the four equating methods and the population equating results are provided in Figure 3. From this figure, the equating relationships for the two traditional methods are noticeably different from the equating relationships for the two IRT methods. This difference is likely a result of the very different statistical assumptions used in each method (Kolen & Brennan, 2004). FE and CE equating relationships are more similar to each other than they are to those for the two IRT methods. The equating relationships for the two IRT methods are very similar to each other, and the differences between the two relationships for the two IRT methods are less than .25 score point across most of the score range (10 to 65), with the largest differences occurring at the very low and very high scores. Also, the IRT equating methods produced smoother equivalents than the FE and CE methods. However, all equating methods produced noticeably different equating results from the population equating relationship.

Comparison of SEE. The *SEE* values for the four equating methods are depicted in Figure 4. First, the *SEE* values for the FE method are smaller than those for the CE method at most score points, which is consistent with the findings of other researchers (e.g., Wang, Lee, Brennan, & Kolen, 2008). Second, the IRT observed score method produces smaller *SEE* values than the IRT true score method and the FE method at most score points.

Overall, *MSEE* provided in Table 3 was smallest for the IRT observed score method (*MSEE*=.3543), followed by the IRT true score method (.3964), which is slightly smaller than that of the FE method (*MSEE*=.4043), and the CE method produces the largest *MSEE* (.4708) among the methods.

Comparison of Bias and RMSE. *Bias* and *RMSE* for the four equating methods are provided in Figures 5 and 6. Over most of the score points, the following was found: 1) The *Bias* indices for the two IRT equating methods are smaller than those for the two traditional equating

methods; 2) the *Bias* indices for the IRT observed score equating methods are slightly smaller than those of the IRT true score method; 3) the *Bias* indices for the frequency estimation equipercentile equating method are smaller than those for the chained equipercentile equating method at low (<15) and high (>50) score ranges. Similar trends can be seen for the *RMSE* indices.

ABIAS and *ARMSE* of the four equating methods for the History test are also provided in Table 3. The results indicate that the two IRT equating methods produce smaller *ABIAS* and *ARMSE* than the two traditional equating methods. Also, the IRT observed score equating method produces smaller *ABIAS* and *ARMSE* than the IRT true score equating method and the CE method has smaller *ABIAS* and *ARMSE* than the FE method.

Biology Test

Comparison of the equating relationships for the sample data. Figure 7 provides the equating relationships for the four equating methods and the population equating results for the Biology test. The equating relationships for the four equating methods are very similar to each other and to the population equating results. IRT true score equating behaves slightly differently from the other three at the score points from 65 to 80. Even in that range, the largest difference among the four methods is less than one score point. The two IRT methods are very similar except at the very low score range (<10) and the range of 65 to 80. Still, the IRT equating methods produced smoother equivalents than FE method and the CE method.

Comparison of the SEE. The *SEE* values for the four equating methods for the Biology test are depicted in Figure 8. Results are similar to those for the History test. First, the *SEE* values for the FE method are smaller than those for the CE method at most score points. Second, the IRT observed score equating method produces smaller *SEE* values than the IRT true score method and the FE method at most score points. Also, the IRT true score equating method had smaller *SEE* values than the FE method over most of the score points.

In Table 4, *MSEE* for the IRT observed score method is .3638, which is the smallest among the four equating methods. *MSEE* for the IRT true score equating method (.3921) is slightly larger than *MSEE* of the IRT observed score equating method and is noticeably smaller than *MSEE* for the two traditional equating methods. The CE method produces the largest *MSEE*.

Comparison of Bias and RMSE. The *Bias* and *RMSE* indices for the four equating methods for the Biology test are provided in Figures 9 and 10. Unlike the results for the History

test, the results for the Biology test are inconclusive from these two figures except that over most of the score points, the IRT observed score method has smaller biases and *RMSE* indices than the IRT true score equating method. No clear conclusion can be made about superiority among the IRT observed score equating, FE, and CE methods. It is difficult to say which method produces smaller biases or *RMSE* indices among these three methods by inspecting these figures.

With respect to *ABIAS* and *ARMSE* summarized in Table 4, the IRT observed score equating method produces smaller *ABIAS* and *ARMSE* than the IRT true score equating method. The FE method has similar *ABIAS* and smaller *ARMSE* compared to those for the CE method. Overall, the IRT true score equating method produces the largest *ABIAS* and *ARMSE* among the four equating methods. No method produces both the smallest *ABIAS* and the smallest *ARMSE*.

Discussion

In this paper, a population equipercentile equating relationship was created by splitting one form into two pseudo forms with some items in common so that the data collection design for the population is a single group design. Then nonequivalent groups were sampled from the population and were administered the two pseudo forms. Using such a procedure, a common item nonequivalent groups design for the two pseudo forms was created. Two empirical data sets were used — one is from the History test and the other is from the Biology test.

Based on this study, the equating relationships for the four equating methods are quite different for the History test and very similar for the Biology test. The results for the standard errors of equating are consistent across the two tests. The two IRT equating methods have smaller standard errors of equating than the two traditional equating methods. More specially, the standard errors of equating for the IRT observed score equating method are slightly smaller than those for the IRT true score equating method, which has smaller standard errors of equating than the frequency estimation equipercentile method. The chained equipercentile method has the largest standard errors of equating.

The results for the bias and root mean square error indices of equating are different across the two tests. For the History test, the following results are found over most of the score points: 1) The bias and root mean square errors for the IRT equating methods are smaller than those for the traditional equating methods; 2) the IRT observed score equating method has smaller bias and root mean square error indices than the IRT true score equating method; 3) the bias and root mean square error of equating indices for the frequency estimation equipercentile equating

method are smaller than those for the chained equipercentile equating method at low (<15) and high (>50) score ranges. For the Biology test, the IRT observed score equating method has smaller bias and root mean square error indices than the IRT true score equating method. No clear conclusions can be made among the IRT observed score equating, frequency estimation equipercentile method, and the chained equipercentile method. It is difficult to say which method produces smaller biases or *RMSE* indices among these three methods for the Biology test. However, the *ARMSE* for the IRT observed score equating method is the smallest among all four equating methods.

Limitations and Future Work

In this study, log-linear presmoothing with C=6 for both forms was used to smooth the population distribution and then the equipercentile equating was used to find the equating relationship for the population data and these equivalents were treated as the population equivalents. The application of equipercentile equating on the population data might favor the equating methods involving equipercentile equating, which included frequency estimation equipercentile equating, chained equipercentile equating, and IRT observed score equating. It would be interesting to investigate the sensitivity of the results to choice of the single group equating criterion that could include the unsmoothed equipercentile equating relationship, the smoothed equipercentile equating relationship using log-linear presmoothing as was done here, the IRT true score equating relationship, and the IRT observed score relationship.

Only one set of pseudo test forms was created for both History test and Biology test, and the results based on the pseudo test forms were different for History test and the Biology test. In order to make a clearer comparison among these four equating methods, more pseudo test forms should be investigated.

Cubic spline postsmoothing with S=.1 was used for the frequency estimation equipercentile equating method and chained equipercentile equating method. Larger S values (for example, S=.3 or .5) should be investigated to compare the traditional equating methods with the IRT equating methods since the choice of a larger S value will likely reduce more the standard errors of equating for the equipercentile equating methods.

In this study, the common item set contained only multiple choice items due to the test security purposes. Therefore, the use of constructed response items in the common item set should be studied for the mixed-format tests.

The effect sizes of the common item means for the examinee groups taking the two forms used in this study are very small (.1391 for the History test and .0732 for the Biology test). Larger effect sizes should be considered in the future studies. In addition, equating situations with mixed-format tests of different lengths, with different score distributions, and from different content areas should be studied.

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.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Cho, Y. W. (2007). *Comparison of bootstrap standard errors of equating using IRT and equipercentile methods with polytomously-scored items under the common-item nonequivalent-groups design.* Unpublished Ph.D. Dissertation, The University of Iowa.

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.

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32,* 334-347.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., Martin, K. (2006). *An evaluation of the kernel equating method: a special study with pseudotests constructed from real test data.* (Research Report RR-06-02). Princeton, NJ: Educational Testing Service.

Harris, D. J. (1991). A comparison of Angoff's design I and design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28,* 221-235.

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.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

Kim, S., Walker, M. E., & McHale, F. (2010). Investigating the Effectiveness of Equating Designs for Constructed-Response Tests in Large-Scale Assessments. *Journal of Educational Measurement, 47,* 186-201.

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

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement,* 9, 209-223.

Kolen, M. J. (2004). POLYEQUATE [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

Liou, M., Cheng, P. E., & Johnson, E. J. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21,* 349-369.

Liu, Y. M., Schulz, E. M., & Yu, L. (2009). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33,* 257-278.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

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, NJ: Educational Testing Service.

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.

Sinharay, S., & Holland, P. W. (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.

Sinharay, S., & Holland, P. W. (2008). *The missing data assumptions of the nonequivalent groups with anchor test (NEAT) design and their implications for test equating*. (Technical Report RR-09-16). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2010). *A new approach to comparing several equating methods in the context of the NEAT design*. *Journal of Educational Measurement, 47*, 261-285.

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

Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG (Version 7.03) [Computer software]. Lincolnwood, IL: Scientific Software International.

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. Y., 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.

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

Zeng, L. J., Hanson, B. A., & Kolen, M. J. (1994). Standard errors of a chain of linear equating. *Applied Psychological Measurement. 18,* 369-378.

Table 1

Mean and Standard Deviation for the Two Population Groups

Test	Form	mean	sd	skewness	kurtosis
History	X	35.7534	14.4198	.1345	-.7315
	Y	35.9160	14.2458	.0559	-.7247
Biology	X	55.4108	17.2098	-.2704	-.5373
	Y	50.8828	16.7930	-.1414	-.4819

Table 2

Descriptive Statistics of Pseudo-tests for the Two Sample Groups

Test	Group	Score	mean	sd	correlation	Effect size
History	1	X	36.8750	14.1184	.8104	.1391
	1	V	7.4277	2.9896		
	2	Y	34.8167	14.3391	.8198	
	2	V	7.0133	2.9659		
Biology	1	X	55.9930	16.8904	.8538	.0732
	1	V	12.8190	3.5320		
	2	Y	50.3633	16.7523	.8554	
	2	V	12.5630	3.5778		

Table 3

Mean Standard Error of Equating, Average Bias, and Average Root Mean Square Error of the Four Equating Methods for the History Test

Equating	IRT-True	IRT-Observed	FE-*S*=.1	CE-*S*=.1
MSEE	.3964	.3543	.4043	.4708
ABIAS	.7341	.6555	1.2985	1.1869
ARMSE	.8343	.7452	1.3600	1.2769

Table 4

Mean Standard Error of Equating of the Four Equating Methods for the Biology Test

Equating	IRT-True	IRT-Observed	FE-*S*=.1	CE-*S*=.1
MSEE	.3921	.3638	.4495	.5189
ABIAS	.6776	.4629	.4099	.3905
ARMSE	.7829	.5888	.6084	.6495

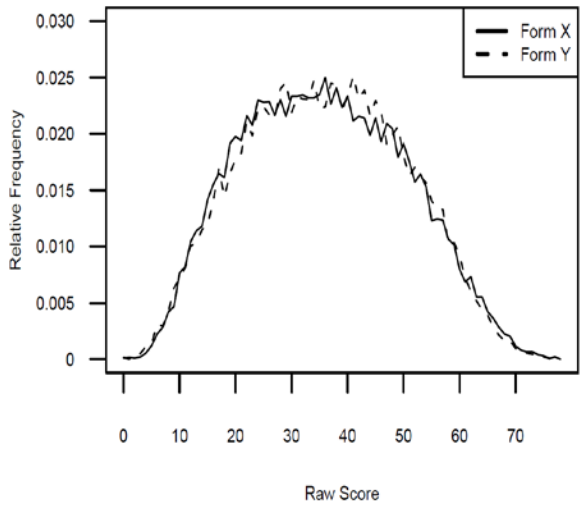

Figure 1. Relative frequency distributions for the population taking the History test.

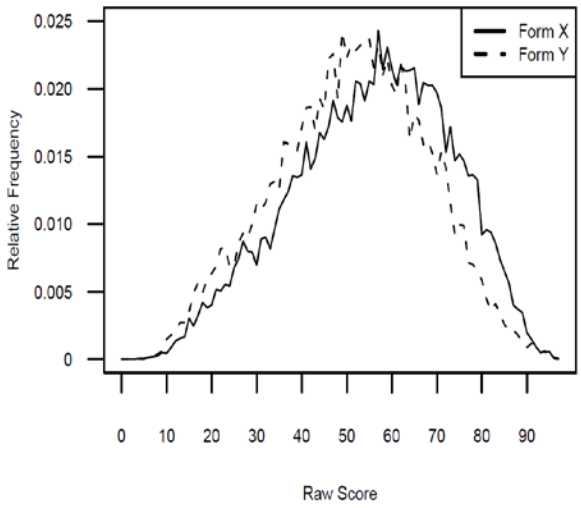

Figure 2. Relative frequency distributions for the population taking the Biology test.

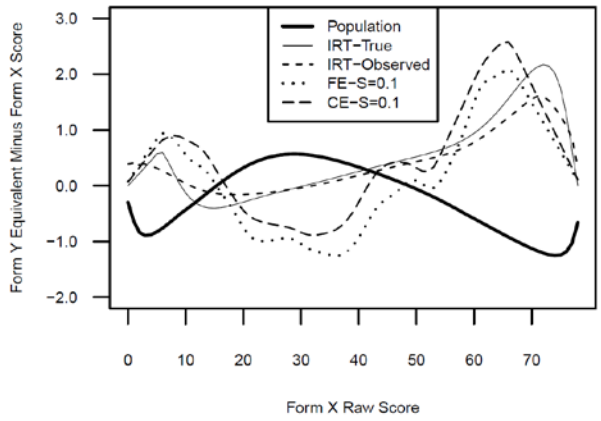

Figure 3. Estimated equating relationships of the four equating methods for the History test.

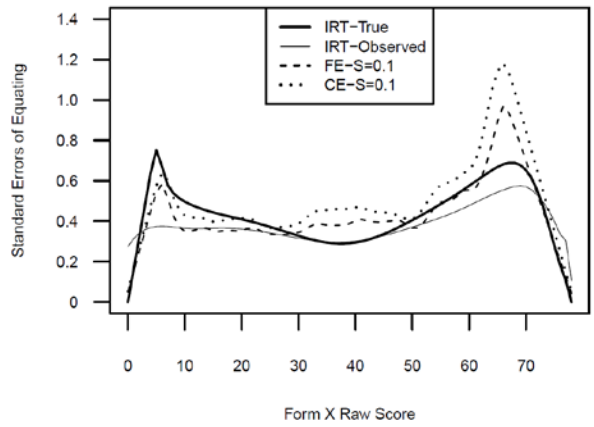

Figure 4. Comparison of the standard errors of equating among the four methods for the History test.

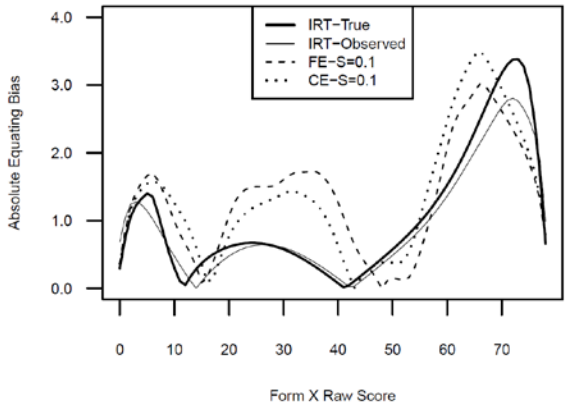

Figure 5. Comparison of equating bias among the four equating methods for the History test.

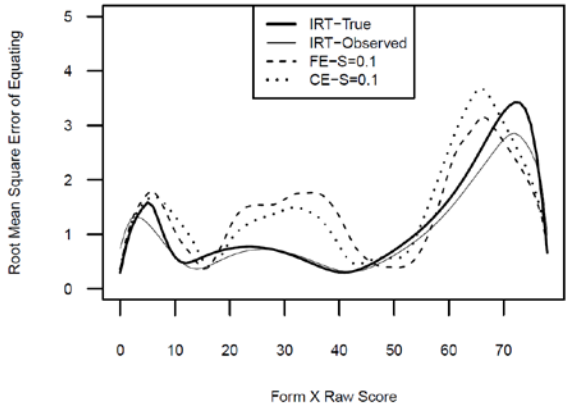

Figure 6. Comparison of root mean square error among the four equating methods for the History test.

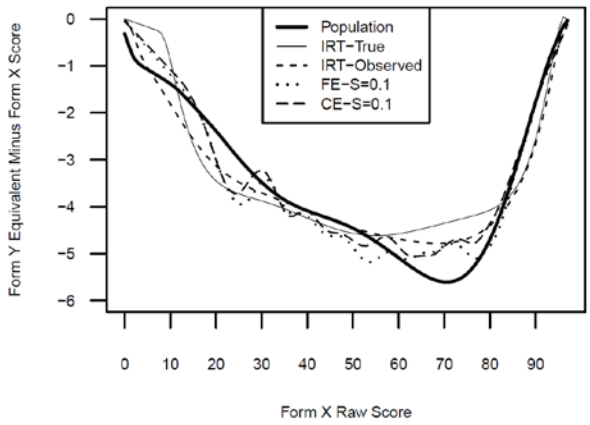

Figure 7. Estimated equating relationships of the four equating methods for the Biology test.

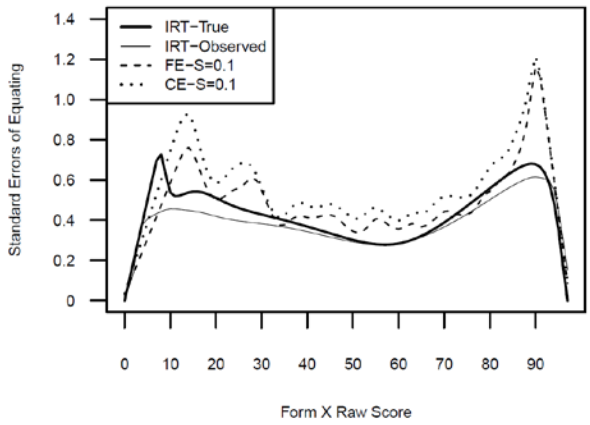

Figure 8. Comparison of the standard errors of equating among the four methods for the Biology test.

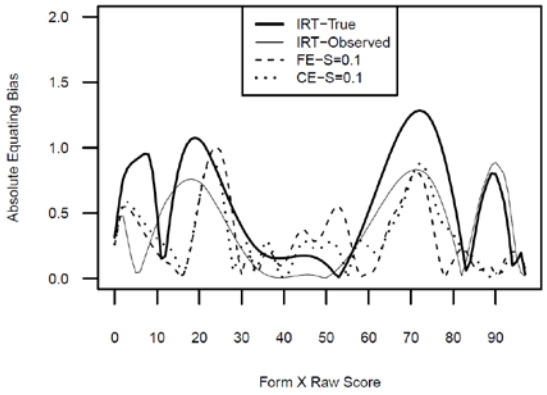

Figure 9. Comparison of equating bias among the four equating methods for the Biology test.

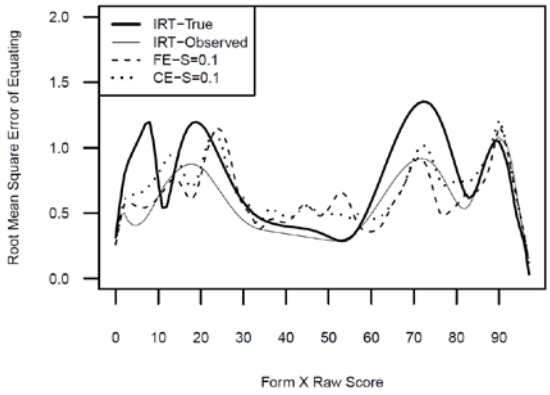

Figure 10. Comparison of root mean square error among the four equating methods for the Biology test.

Chapter 5: Equating Mixed-Format Tests with Format Representative and Non-Representative Common Items

Sarah L. Hagge and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

Mixed-format tests containing multiple-choice and constructed-response items are widely used on educational tests. Such tests combine the broad content coverage and efficient scoring of multiple-choice items with the assessment of higher-order thinking skills thought to be provided by constructed-response items. However, the combination of both item formats on a single test complicates the use of psychometric procedures. The purpose of this study was to examine how characteristics of mixed-format tests and composition of the common-item set impact accuracy of equating results in the common-item nonequivalent groups design. Pseudo-test forms assembled from a Spanish Language mixed-format test from the Advanced Placement Examination program were considered for this study. Four factors of investigation were considered: (1) difference in proficiency between old and new form groups of examinees, (2) relative difficulty of multiple-choice and constructed-response items, (3) format representativeness of the common-item set, and (4) equating methods. For each study condition, frequency estimation equipercentile, chained equipercentile, item response theory (IRT) true score, and IRT observed score equating methods were considered. There were four main findings from the pseudo-test form analyses: (1) As the difference in proficiency between old and new form groups of examinees increased, bias also tended to increase; (2) relative to the criterion equating relationship for a given equating method, increases in bias were typically largest for frequency estimation and smallest for the IRT equating methods; (3) standard errors of equating tended to be smallest for IRT observed score equating and largest for chained equipercentile equating; and (4) bias tended to be smaller for the common-item composition containing both multiple-choice and constructed-response item formats.

Equating Mixed-Format Tests with Format Representative and Non-Representative Common Items

Many testing programs now widely use both multiple-choice (MC) and constructed-response (CR)[1] item formats in mixed-format tests. By using both formats in the same test, tests combine the high reliability and broad content coverage afforded by MC items while using CR items to closely represent real-life situations and to possibly invoke complex levels of thought. Although mixed-format tests combine the positive aspects of both item formats, they also combine the challenges. The purpose of the current study is to understand the impact that the inclusion of CR items has on current psychometric methods. Specifically, this study examines how the characteristics of mixed-format tests might impact equating and aims to explore test characteristics that may lead to satisfactory equating. The potential problems associated with equating mixed-format tests, and the potential benefit to operational testing programs of knowing which test characteristics are likely to be conducive to adequate equating, motivated this study. Specifically, the current study addresses the following questions:

1. What is the impact on equated scores when examinees on one mixed-format test form are higher in proficiency than examinees on the other mixed-format test form?
2. When one type of item format (i.e., MC or CR) is relatively more difficult for examinees taking one form as compared to examinees taking another form, how are the resulting equated scores impacted?
3. How does the composition of the common items impact equated scores?

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). The possible answer choices include one correct answer and a number of incorrect answer choices (Ebel & Frisbie, 1991). Many of the strengths of MC items lie in efficiency of administration and scoring. However, MC items have been criticized for possibly not assessing higher-order thinking skills and curricular overemphasis on MC specific test-taking strategies (Ferrara & DeMauro, 2006).

[1] In this chapter, the phrase "constructed-response" is used instead of "free-response," which is used in all other chapters.

CR items require examinees to produce an answer or product without existing answer choices. CR items can be classified into many different categories, such as short answer, products, performances, completion, or construction (Bennett, 1993; Ferrara & DeMauro, 2006). Some of the strengths of CR 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, CR items are more expensive to score, easier to memorize, and typically have lower reliability per unit of testing time compared to MC items. 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 CR items appears to have been found with mixed-format tests that contain both MC and CR items. By including each item format on a test, some of the weaknesses of each item format are mitigated by the strengths of the other. However, the combination of both item formats on the same test introduces other potential problems for existing psychometric methodologies, such as multidimensionality or how to appropriately weight MC and CR items to create composite scores. Of particular interest for this study is the way in which characteristics of mixed-format tests impact equating.

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). Different forms of the same test are rarely, if ever, created perfectly parallel. Consequently, test forms almost always differ in difficulty, necessitating the use of equating. Without equating, if test forms differ in difficulty, examinees from one administration may be unfairly advantaged or disadvantaged relative to examinees taking a test form from a different test administration. The focus of the current study is on the common-item nonequivalent groups (CINEG) equating 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 each of two test forms. The common items allow for separation of differences in examinee score distributions across the two test forms into differences in form difficulty and differences in 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.

Potential Problems for Equating Mixed-Format Tests in the CINEG Design

One of the crucial components of the CINEG design is the careful development of common-item sets that accurately reflect the total test. For MC tests, Kolen and Brennan (2004) discussed desirable characteristics of common-item sets. They indicated that the common items should be a "mini version" of the total test. The common items should be "proportionally representative of the total test forms in content and statistical characteristics" (Kolen & Brennan, 2004, p. 19). Additionally, common items should remain unchanged from the old to the new form and be placed in the same position on both forms.

The extent to which the guidelines for common items on MC-only tests generalize to mixed-format tests is not well known. Presumably, the same guidelines for MC-only tests would also hold for mixed-format tests. However, the additional consideration of item format must be taken into account. The rationale for the inclusion of CR items in tests is that they measure content or processes that cannot be adequately measured by MC items. Therefore, it is reasonable to assume that common items should also be representative of the total test in terms of format. However, including CR items as common items might not be feasible or even desirable (Kirkpatrick, 2005). There are typically only a limited number of CR items to select as common items. Additionally, security of the CR items could be compromised, because it is potentially easier to memorize one of only a few CR items. Further, CR items are typically scored by at least one human rater, introducing another source of error for CR items. Rater leniency may not be consistent across years. Consequently, for each administration, CR items used as common items would need to be rescored for a sample of examinees, a process known as trend scoring, resulting in additional costs to the testing program.

The challenges in using CR items as common items for mixed-format tests have resulted in many testing programs electing to use MC-only common items for mixed-format tests. Therefore, it is important to understand which characteristics of a mixed-format test might impact equating results with MC-only common items.

Overview of Relevant Research

Research on equating has taken into account many factors, including equating methods and test characteristics, examinee proficiency, and composition of the common-item set.

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

Equivalence of constructs measured by MC and CR items. One factor that may impact equating is the extent to which MC and CR items are measuring equivalent constructs on a given test. The extent to which MC and CR items measure equivalent constructs may differ according to the subject area as well as format of the CR item. Some evidence has suggested that in the writing domain, MC and CR items may not be measuring equivalent constructs. In the quantitative domain, there is evidence to suggest that MC and CR items may measure similar constructs (Traub, 1993). However, factor analyses have indicated the presence of a possible weak format factor, even in the quantitative domain (Sykes, Hou, Hanson, & Wang, 2002; Thissen, Wainer, & Wang, 1994). Other studies have suggested that when MC and CR items are designed to measure equivalent constructs, they perform similarly (Rodriguez, 2003).

Test, common item, and examinee characteristics. Literature has indicated that characteristics of the test and common-item set impact equating results, although results were often dependent on differences in examinee proficiency across test forms. For MC-only tests, the spread of difficulty of the items in the common-item set did not result in large practical differences in equating (Sinharay & Holland, 2007). However, a shift in the mean difficulty on a mixed-format test was found to provide less accurate equating results (Cao, 2008). Increasing the number of items or score points on a CR test appeared to increase equating accuracy, although larger increases were seen with increases in the number of items (Fitzpatrick & Yen, 2001).

A number of research studies have also investigated the impact on equating results of using an MC-only common-item set or format representative common-item set. In some cases, MC-only common items resulted in considerable bias (Walker & Kim, 2009), but studies have suggested that higher correlations between MC and CR items, smaller group differences, and higher MC to CR point ratios may result in less biased equating relationships (Cao, 2008; Kirkpatrick, 2005; Lee, Hagge, He, Kolen, & Wang, 2010; Wang, Lee, Brennan, & Kolen, 2008; Wu, Huang, Hu, & Harris, 2009; Tan, Kim, Paek, & Xiang, 2009). Other research suggests common items containing both MC and CR items may result in the most accurate equating when a test is multidimensional or group proficiency differs across item formats (Cao, 2008; Kirkpatrick, 2005). However, use of trend scoring may be recommended when CR items are

included in the common-item set (Kim, Walker, & McHale, 2008; Tate, 2003). Using only the dominant item type included on the test may also lead to reasonable equating results in some cases (Kim & Lee, 2006).

The potential problems associated with equating mixed-format tests, and the potential benefit to operational testing programs of knowing which test characteristics are likely to be conducive to adequate equating motivated this study. This study seeks to investigate when use of MC-only common items yields plausible equating results and whether the addition of CR items increases the accuracy of equating relationships.

Methodology

The methodology is divided into four sections. The first two sections describe the original operational test form used for the study. The last two sections describe the construction of pseudo-test forms and factors of investigation for the pseudo-test forms.

Selection of Tests

Data for this study were from a Spanish Language College Board Advanced Placement (AP) test. Although an AP test was used for analyses in this study, the exam was manipulated in order to investigate how equating methods, test characteristics, and differences in examinee group proficiency affect equating results for mixed-format tests. The test was modified in such a way that the characteristics of the test 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. One Spanish Language test form (2004) was selected for analyses for the current study because of the variety of CR prompts included in the exam. The Spanish Language test addressed listening, reading, writing, and speaking skills. The MC items measured listening and reading comprehension. CR 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 test form contained 90 MC items and 27 CR items. The 27 CR items were spread across three different formats. One item-format was fill-in-the-blank and consisted of 20 items worth one point each. The second item format was speaking prompts consisting of five items worth four points each. The third item format, written communication and essay, consisted of two items worth nine points each. In total, there were 58 CR points.

Data Preparation

For this study, number-correct scoring was used for the MC items. However, formula scoring was used for the MC items on the original operational test forms. Examinees were advised that they would be penalized for incorrect answers on the MC items. Consequently, some examinees had missing responses for a large number of MC items. In order to use number-correct scoring and use item response theory (IRT) equating methods, examinees completing fewer than 80% of the MC items were removed. For the remaining examinees, incorrect responses were coded as 0, correct responses were coded as 1, and for missing responses, imputation was used to replace each missing response with a correct or incorrect response. 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 1. In Table 1, means, standard deviations, skewness, and kurtosis are given for MC, CR, common-item (CI), and composite (CO) scores which are the sum of the MC and CR scores. For MC, CR, and CO, α is also given. For CI, the correlation between composite and common-item scores is given (CO Corr.).

Operationally, weights used for AP exams are very 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. Although the choice of weighting may impact equating results, weighting was not considered as a factor of investigation in this study. Summed scores were used. That is, each MC or CR point was worth one point and was not differentially weighted.

Dimensionality Assessment

In order to evaluate whether the IRT assumption of unidimensionality held, a dimensionality assessment was conducted. For this study, disattenuated correlations between MC and CR scores were considered using Cronbach's alpha, and a principal components analysis (PCA) was conducted using tetrachoric and polychoric correlations among all individual items. Disattenuated MC and CR correlations near 1.00 indicated the MC and CR sections of the test were measuring essentially the same dimensions. For the PCA, four rules of thumb were used to determine whether a test was sufficiently unidimensional for IRT analyses. The first guideline

was to retain only those 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 what might be considered 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 (Divgi, 1980; Hattie, 1985; Jiao, 2004; Lord, 1980). If this ratio was larger than 3, then the first principal component was considered strong relative to the other principal components, indicating a unidimensional tendency. 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.

Construction of Pseudo-Test Forms

In this study, data on pseudo-test forms were used to examine the impact of test and common-item composition on equating, which could not be investigated through operational test form analyses. Pseudo-test forms were created by splitting one test form into two new test forms. The Spanish Language 2004 operational test form contained 90 MC items and 27 CR items. Old and new pseudo-test forms were created from the original operational test form. The new and old pseudo-test forms each contained 21 unique MC items and seven unique CR items. In addition, the old and new pseudo-test forms contained 46 MC items and 13 CR items in common with each other. All of the items contained in the new and old pseudo-test forms were originally on the operational Spanish Language 2004 form. It is important to note that although the old and new Spanish Language pseudo-test forms contained 46 of the same MC items and 13 of the same CR items, only a subset of these items were actually used as common items for analyses, which is discussed in a later section.

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 CR item and the total scores were calculated as discrimination.

Factors of Investigation

Four factors of investigation were considered for the pseudo-test form analyses: examinee common-item effect sizes, MC and CR relative difficulty, composition of the common-item set, and equating methods.

Examinee common-item effect sizes. The examinees taking the old and new pseudo-test forms were the same examinees. Consequently, 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. The demographic variables used in this study were ethnicity and parental education level. It was determined that examinees belonging to different demographic groups performed differently across common-items, CR items, and MC items. Therefore, by oversampling certain demographic groups using an iterative procedure, different common-item effect sizes could be obtained across old and new forms.

Common-item effect sizes (CI ES) indicated the standardized amount mean common-item scores of the examinees differed across old and new test forms. Four levels of CI ES were chosen for this study: .00, .20, .40, and .60. The lower two levels were selected to represent differences in mean proficiency that might be seen in practice. The highest levels were selected to represent a large CI ES. Kolen and Brennan (2004) stated that differences of .30 standard deviation units or more could result in large differences across equating methods. ES were calculated for new form minus old form common-item scores, as shown in Equation 1,

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

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

$s_2^2 = \text{Variance of scores for old form.}$

MC and CR relative difficulty. If MC and CR items measure different constructs, it is plausible that examinee proficiency may differ across item formats. Such a difference might occur, for example, if instruction for one group focused more on the construct assessed by the CR items for one group than for the other group. When the common items contain only items representative of the MC items, there is an implied assumption that examinees will perform similarly on both MC and CR items. If examinee proficiency differs across formats, having only MC items in the common-item set may be problematic for equating. Research Question Two addressed this possibility. A procedure similar to the procedure for producing group differences in the common-item mean scores was used to create differences in the relative difficulty of MC and CR items. The same formula was used to calculate effect sizes. However, effect sizes were calculated for new form scores minus old form scores for both MC and CR scores, separately, on the entire form (not just on the common items). In order to determine the difference in relative difficulty across old and new forms for MC and CR, the difference in effect sizes for MC and CR was calculated. This effect size is referred to as the MC-CR ES. Two target levels of MC-CR ES were selected: .00 and .25. These two levels were combined with the levels of CI ES to create eight combinations of CI ES and MC-CR ES for the pseudo-test forms. An iterative sampling process was used to sample examinees from demographic groups to achieve the desired effect sizes (see Hagge, 2010). For this study, a sample size of 1,900 was selected. This was the largest sample size that could be obtained for the most extreme combination of CI and MC-CR ES. It is important to note that, because of limitations of sampling real data, the exact effect sizes that were desired could not always be achieved.

Common-item composition. Content representation of the common-item set was not considered in this study, because a number of factors limited the feasibility of developing content representative common items. In order to ensure reasonable content representation, common-items from the operational test administrations were used whenever possible. One common-item set length was considered for this study. A length of approximately 30% of the total points on the test was chosen for the length. One typical rule-of-thumb for common items is that the common items should be at least 20% of the length of the total test (Kolen & Brennan, 2004). Many of the CR items on the tests in this study had large point values. Restricting the common-item set to only 20% of the length of the total test limited the extent to which CR items could be used in

common items. Therefore, a longer common-item length was selected in order to be able to include an adequate number of points allocated to MC and CR items.

Two levels of format representativeness were used in this study: **N**o **CR** representation (NCR) and **F**ull **CR** representation (FCR). No CR representation means that the common items contained only MC items. Full CR representation means that the items in the common-item set were proportionally representative of the formats on the total test. That is, the common-item set contained 30% of the total points allocated to MC and 30% of the total points allocated to CR. In order to hold test characteristics constant, the items on each test form were the same across common-item conditions. The only aspect that varied was which items were treated as common items. Consequently, both the old and new pseudo-test forms contained some items in common that were not treated as common for all of the common-item set conditions. Additionally, it is important to note that the majority of the MC common items were common items on the original operational test forms.

Table 2 contains information regarding the composition of the pseudo-test forms and common-item sets. The first column indicates old or new pseudo-test forms and the second column contains the item type. For example, total test refers to the items on the total test, NCR refers to MC-only common items composed to be a minitest, FCR refers to combined MC and CR common items, and Full CI refers to all items in common between the old and new pseudo-test forms. Even though equating was not conducted using Full CI, it is reported because sampling was based on examinee scores on Full CI. The next two columns contain the number of MC items (number of MC points in parentheses) and number of CR items (number of CR points in parentheses), respectively. The last four columns of data in Table 2 contain the mean and standard deviation of difficulty and discrimination. For example, the NCR set contained 33 MC items worth 33 points. The FCR set contained 21 MC items worth 21 points and six CR items worth a total of 12 points. As intended, mean difficulty for NCR and FCR were similar to each other and to the mean difficulty for the total test, as shown in the fifth column. Standard deviations of difficulty were also similar to the standard deviation of difficulty for the total test (within approximately .03). Mean discrimination for the common-item sets were lower than for the total test.

Equating methods. Equating was conducted for the CINEG design using four equating methods: frequency estimation (FE), chained equipercentile equating (CE), IRT true-score

equating (TS) and IRT observed-score equating (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 .1 was chosen based on previous research conducted on the AP exams. For FE, the new form population received a weight of one to create synthetic populations. The equating process was replicated using 500 bootstrap replications. It is important to note that the sample of examinees was a bootstrap sample, meaning that examinees were resampled, with replacement, from the sample of 1,900 examinees.

Before conducting equating for the IRT methods, item parameter estimates and estimates of the distribution of proficiency were obtained using PARSCALE (Muraki & Bock, 2003). The 3PL model was used for the MC items, and the GPCM was used to estimate parameters for the CR 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 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 bootstrap replications. Table 3 contains a summary of the factors of investigation for the pseudo-test form analyses.

The process for conducting bootstrap replications for IRT equating methods was complex and involved a series of computer software packages. It was impossible to check all of the output for accuracy on the individual replications. Therefore, a number of steps were taken in an attempt to ensure that the IRT analyses were accurate. When PARSCALE could not estimate a pseudo-chance item parameter, as evidenced by 0 for the item parameter and standard error estimates, that item parameter was fixed to 0 in the command file. However, for a given test form, the same control card was used across all bootstrap replications. Additional checks were also built into the IRT standard error of equating computer code. These checks included removing replications when PARSCALE did not converge and removing replications for which the equating results seemed unreasonable. Most commonly, equating results were removed when a problem occurred during equating which resulted in the majority of the new form raw scores being equated to the highest possible old form raw score. The number of replications removed for this reason was less than 1% across all analyses and was typically only one or two replications.

Evaluation

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 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. To evaluate the results from the pseudo-test form analyses at each raw score point, conditional 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.

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

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

$$RMSE_i = \sqrt{Bias_i^2 + CSE_i^2}. \quad (4)$$

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, $\hat{e}_j(x_i)$ is the old form equivalent of a new form raw score for a study condition equating relationship, and $\bar{\hat{e}}(x_i)$ is the average of $\hat{e}_j(x_i)$ over replications. 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. Plots of conditional bias, RMSE, and CSE were created to graphically represent differences between the criterion and study condition equating relationships.

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.

$$WARMSB = \sqrt{\sum_i w_i Bias_i^2}\,. \quad (5)$$

$$WASE = \sqrt{\sum_i w_i CSE_i^2}\,. \quad (6)$$

$$WARMSE = \sqrt{\sum_i w_i RMSE_i^2}\,. \quad (7)$$

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. Consequently, it was desirable to weight those score points so that they received less weight relative to scores where a large number of examinees scored.

Results

The results section is arranged into two sections: dimensionality assessment and pseudo-test forms. The first section provides results for the dimensionality analyses, and the second section contains results for the pseudo-test forms.

Dimensionality Assessment

Correlations for the Spanish Language pseudo-test forms and Spanish Language 2004 operational test form are presented in Table 4. The disattenuated correlations were below 1, suggesting that the MC and CR items were measuring somewhat different constructs. Correlations for the pseudo-test forms were similar in magnitude to the operational test form. For the original Spanish Language 2004 test form, thirteen eigenvalues were one or greater than one. By examining the scree plot in Figure 1, it is evident that the first and second eigenvalues were substantially larger than the other eigenvalues. The two large eigenvalues relative to the other eigenvalues suggest potentially two stronger dimensions. Additionally, the eigenvalues did not level off until after the fourth eigenvalue, suggesting the possibility of two weaker dimensions. However, the first principal component accounted for 27% of the total variance. Additionally, the ratio of the difference between the first and second eigenvalues and the difference between the second and third eigenvalues was approximately 4.5. The conclusions from the four criteria for evaluating the PCA results along with the disattenuated MC and CR correlations suggest it is

plausible that more than one dimension may have existed for Spanish Language, although three of the four criteria for the PCA indicated the test was satisfactorily unidimensional for IRT analyses.

Pseudo-Test Form Results

Results for the Spanish Language pseudo-test forms are presented in the following subsections: descriptive statistics, equating relationships, conditional bias, and overall summary statistics. Throughout the pseudo-test forms results section, when different compositions of common items are considered (i.e., NCR or FCR), they are referred to as common-item sets. The abbreviation, CI, is used to indicate different CI ES only.

Descriptive statistics. Descriptive statistics for the old and new Spanish Language pseudo-test forms are provided in Tables 5 and 6, respectively. Descriptive statistics for the common items are especially important to consider. In both tables, NCR refers to the minitest common-item set that contains only MC items and FCR refers to the common-item set containing both MC and CR items. Full CI refers to all items in common between the two test forms. Means for FCR were generally higher than for NCR. Standard deviations were generally smaller for FCR. It is important to note that the skewness varied substantially across sampling conditions. The rows labeled "CO Corr." contain correlations between CI and CO scores. The FCR common-item set always resulted in the largest correlation between CI and CO scores.

Table 7 contains effect sizes based on new form minus old form scores; therefore, negative scores indicate that means on the new form were lower than means on the old form. The first row contains effect sizes for CO scores. The second block of values contains effect sizes for NCR, FCR, and Full CI. The MC, CR, and MC-CR effect sizes are contained in the last block. It is important to note that the MC-CR effect size for the single group was -.270. Therefore, for the MC-CR .00 sampling conditions, the MC-CR effect sizes were similar to this value. All of the MC-CR .00 effect sizes were within .02 standard deviation units of -.270. The MC-CR effect size for the MC-CR .25 sampling conditions differed from -.270 by approximately .25 standard deviation units. The effect sizes for the MC-CR .25 sampling conditions were approximately 0. Although the magnitude of the effect sizes and labels is counterintuitive, remember that MC-CR .00 represents the sampling conditions with the MC-CR effect size most similar to the criterion. MC-CR .25 represents the sampling conditions with the MC-CR effect sizes most different from the criterion. Also, recall that 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 .035 of the target effect sizes.

Equating relationships. Equating relationships discussed in this section were calculated for the original samples of 1,900 examinees per form, and not over the bootstrap replications. As previously described, the criterion equating relationships for Spanish Language pseudo-test forms were based on a single-group equating. For both the FE and CE methods, the criterion was an equipercentile equating relationship. For TS and OS, the criterion equating relationships were TS and OS equating relationships, respectively. Figure 2 contains a plot of the three criterion equating relationships. It is evident in Figure 2 that the three single-group equating relationships were very similar across the score scale.

Figures 3 through 5 contain plots of the equating relationships. Figure 3 contains plots for FE and Figure 4 contains plots for CE. Results for TS and OS were nearly identical, so plots are included only for TS in Figure 5. Each figure contains four plots. The left column contains plots for the MC-CR .00 equating relationships, and the right column contains plots for the MC-CR .25 equating relationships. Each row contains plots for a different common-item set composition. The first row in Figure 3 contains plots for NCR for the FE equating method. The second row in Figure 3 contains plots for FCR for the FE equating method. In each plot, there are five lines, one for each CI ES sampling condition (i.e., CI .00, CI .20, CI .40, and CI .60) and one for the criterion equating relationship. The criterion equating relationship is illustrated by the solid dark black line. The CI ES equating relationships are illustrated by solid (CI .00), dashed (CI .20), dotted (CI .40), and dotted-dashed (CI .60) lines.

Consider Figure 3 for the FE equating method first. The plot on the left of the first row is for NCR for the MC-CR .00 sampling conditions. The difference between the study condition equating relationships and the criterion equating relationship increased as the CI ES increased. It is important to note that the equating relationship for CI .00 MC-CR .00 resulted in slightly lower equated scores than the criterion equating relationship. The plot on the right of the first row is for NCR for the MC-CR .25 sampling conditions. It is evident in this plot that equating relationships for MC-CR .25 resulted in substantially higher equated scores as compared to MC-CR .00. Next, consider the plots for FCR in the last row of Figure 3. By comparing FCR (second row) to NCR (first row), it is evident that the equating relationships for MC-CR .00 (left column) were similar for both common-item sets. However, in the plot for MC-CR .25 (right column), the

equating relationships for FCR were different from the equating relationships for NCR. First, the equating relationships for FCR were more similar to the criterion. Second, the equating relationships for the four CI ES were all similar to each other, especially at scores greater than approximately 60. The pattern of results described for FE in Figure 3 can also be seen in Figure 4 for CE. However, the differences among the equating relationships for different CI ES were smaller for CE as compared to FE. Although the general pattern of equating results was similar for TS (Figure 5) and OS as compared to FE and CE, there were notable differences. In contrast to FE and CE, for the MC-CR .00 sampling conditions (left column), equating relationships differed very little across NCR and FCR for TS and OS.

Conditional bias. Figures 6 and 7 contain plots of conditional bias. Each figure contains four plots of bias. The left column contains plots for the MC-CR .00 equating relationships, and the right column contains plots for the MC-CR .25 equating relationships. Each row contains plots for a different common-item set composition. 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. For brevity, results are presented only for CI .00 and CI .60. Figure 6 contains results for CI .00, and Figure 7 contains results for CI .60.

Consider the plots of conditional bias in Figure 6 for CI .00. By comparing the MC-CR .00 sampling conditions (left column) to the MC-CR .25 sampling conditions (right column), it is evident that the patterns of conditional bias were quite different across the two MC-CR ES levels. For NCR, FE and CE resulted in similar patterns of conditional bias. TS and OS also resulted in similar patterns of conditional bias. However, conditional bias for FE and CE differed from conditional bias for TS and OS by approximately 2 score points. For FCR, all four equating methods resulted in similar conditional bias.

Similar trends can also be seen in Figure 7 for CI .60. Even for CI .60, for both the MC-CR .00 and MC-CR .25 sampling conditions, the four equating methods resulted in similar bias for FCR. For NCR, it is evident than FE and CE resulted in much larger bias than TS or OS.

Overall summary statistics. Tables 8 and 9 contain overall summary statistics calculated over the 500 bootstrap replications. Table 8 contains results for NCR and Table 9 contains results for FCR. Values of WARMSB tended to be highest for FE and CE and lowest for TS and OS. However, for FCR and the MC-CR .25 sampling condition, bias tended to be larger for TS and OS as compared to FE and CE. Values of WARMSB were substantially larger for the MC-CR

.25 sampling conditions as compared to those for MC-CR .00. WARMSB generally increased as the CI ES increased. However, for FCR, values of WARMSB for CI .20 were smaller than values for CI .00. These results were expected given the equating relationships shown in Figures 3 through 6. For TS and OS, values of WARMSB did not exhibit a consistent increasing or decreasing trend.

Comparing the magnitude of WARMSB across NCR and FCR is of particular importance. For the MC-CR .00 sampling conditions, values of WARMSB for FE and CE were typically smaller for FCR as compared to NCR. For the MC-CR .25 sampling conditions, FCR resulted in substantially lower values of WARMSB for FE and CE relative to NCR. However, for TS and OS, values of WARMSB were larger for FCR as compared to NCR by as much as approximately .30 score points. The effect sizes for FCR were much more similar to the composite score effect sizes, especially for large CI ES. Consequently, it appeared that when the relative difficulty of MC and CR items differed across test forms, common items consisting of both MC and CR items more accurately represented the total test.

The second block of data in Tables 8 and 9 contain values for WASE. Values of WASE also tended to be smaller for FCR than for NCR. Across all common-item sets and sampling conditions, OS typically resulted in the smallest values of WASE and CE typically resulted in the largest values of WASE.

The third block of data in Tables 8 and 9 contain values for WARMSE. Values of WARMSE resulted in the same relationships as illustrated by the results for WARMSB. Additionally, values of WARMSE tended to be similar in magnitude to the values of WARMSB.

Summary. As the common-item effect size increased, bias also increased. Of particular importance for the Spanish Language pseudo-test forms is the impact of common-item composition on bias. For the MC-CR .00 sampling conditions, the common-item set containing both multiple-choice and constructed-response items (FCR) resulted in similar, but typically somewhat smaller bias, than the minitest (NCR). However, for the MC-CR .25 sampling conditions, bias for the common-item set containing both multiple-choice and constructed-response items (FCR) was substantially smaller than bias for the minitest (NCR) for FE and CE. Bias did not vary substantially across common-item effect sizes for TS or OS.

Discussion

The purpose of this study was to examine how characteristics of mixed-format tests might impact equating and explore test characteristics that might lead to satisfactory equating with mixed-format tests. The specific factors of investigation in this study included examinee group differences, equating methods, format representativeness of the common-item set, and relative difficulty of MC and CR items. Companion studies to this research have also been conducted on tests in subject areas of English Language and Chemistry (Hagge, 2010). A strength of this study was in the use of pseudo-test methodology. The use of such methodology provided legitimate single group criterion equatings, allowed for the formation of examinee groups that differed by prespecified amounts, and made use of data from operationally tested examinees. This discussion section is divided into four sections: summary of findings, practical implications, limitations, and future research.

Summary of Findings: Research Question One

What is the impact on equated scores when examinees on one mixed-format test form are higher in proficiency, as measured by items in common between test forms, than examinees on the other mixed-format test form? 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 was based on 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. Equating methods generally yielded similar results when examinee groups on the old and new forms were similar in proficiency. As examinee groups differed more in proficiency across old and new test forms, FE generally resulted in larger bias than CE, TS, or OS. TS and OS also typically resulted in smaller bias than CE.

A number of previous research studies examined the impact of group differences on the accuracy of equating. 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., 2010; 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.,

2010; Wang et al., 2008). Research comparing IRT and traditional equating methods for mixed-format tests is less prevalent; therefore, the results from this study comparing traditional and IRT equating methods are somewhat novel. Some previous research has found that TS and CE may both be relatively group invariant (von Davier & Wilson, 2008). However, both Cao (2008) and Kirkpatrick (2005) found that differences in group proficiency impacted equating results in the IRT framework.

Summary of Findings: Research Question Two

When one type of item format (i.e., MC or CR) is relatively more difficult for examinees taking one form as compared to examinees taking another form, how are the resulting equated scores impacted? Results from this study suggest that equating mixed-format tests with only multiple-choice common items may result in larger bias when examinees find certain item formats more difficult relative to other item formats. The MC-CR .00 sampling conditions represented conditions where examinee performance was similar (relative to the criterion) on the MC and CR items. The MC-CR .25 sampling conditions represented conditions where examinee performance was different (relative to the criterion) across the MC and CR items. Therefore, it was expected that bias would be larger for the MC-CR .25 sampling conditions as compared to the MC-CR .00 sampling conditions. As expected, the MC-CR .25 conditions often resulted in larger bias than the MC-CR .00 conditions.

Summary of Findings: Research Question Three

How does the format of the CI impact equated scores? The primary focus of this study was investigation of whether the inclusion of CR items in addition to MC items in a common-item set results in more accurate equating relationships than MC items alone. In general, the common-item set containing both MC and CR items tended to result in smaller values of WARMSB as compared to the MC-only common-item set. However, when the common-item effect size was small (i.e., CI .00 or CI .20), bias was not always smaller for the common-item set containing both item formats. Further, for the TS and OS equating methods, the common-item set containing both item formats did not always reduce bias over the MC-only common-item set. For the MC-CR .25 sampling condition, the common-item set containing both MC and CR items resulted in substantially smaller values of WARMSB as compared to the MC-only common-item set for FE and CE. However, for TS and OS, values of WARMSB were larger for

the common-item set containing both item formats, possibly because of violations of the unidimensionality assumption of the IRT TS and OS methods.

The results from this study regarding format representativeness of the common-item set were mixed, which is consistent with mixed results from previous literature. Kirkpatrick (2005) found that inclusion of a CR item in the common-item set resulted in very small differences among equating relationships. However, when the MC and CR correlation was low or examinee means for MC and CR scores differed, inclusion of a CR item in the common-item set did impact equating results. Cao (2008) also found that when tests were multidimensional, a format representative anchor led to more accurate equating results. However, Kim et al. (2008) found that inclusion of CR items in the common-item set improved equating accuracy over an MC-only common-item set only when the CR items were trend scored. Although it should not be assumed that inclusion of CR items in a common-item set always improves the accuracy of the equating relationship, it is reasonable to assume that in some situations, the equating relationship can be improved by including CR items in the common-item set.

Practical Implications

Composition of the common-item set. Results from this study suggest that using CR items along with MC items in the common-item set may improve equating relationships in certain situations. It is plausible that including CR items in the common-item set may improve equating relationships when examinees perform differently on MC than on CR items for the old and new test forms. It may also be beneficial to include CR items in the common-item set when more than one CR item or item format can be included. However, because the evidence in this study is limited, it is not advisable to draw conclusions about the specific operational settings for which CR items should be included in the common-item set. Further, Kim et al., (2008) suggested that use of CR items without trend scoring in the common-item set would lead to similar results as an MC-only common-item set.

Choice of equating method. When samples of examinees were relatively similar in proficiency across old and new test forms, all of the equating methods resulted in similar equated scores. Therefore, in practice, choice of equating method might be of little consequence when examinee groups are similar in proficiency. However, standard errors of equating were typically larger for CE as compared to FE, TS, or OS. Consequently, when groups are similar in proficiency, FE, TS or OS might be preferred over CE. As the difference in proficiency between

old and new form groups of examinees increased, TS, OS, and even CE were less sensitive to differences than FE. Therefore, for operational equating, it *may* be advisable to consider using TS, OS, or CE when examinee groups differ substantially in proficiency. Further, because only a limited number of effect sizes were considered for this study, it is not possible to determine how large an effect size must be before equating results are no longer reasonable.

Limitations

Resampling operational data. 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 CR items, it was not possible to hold old form means constant across sampling conditions. The lack of control limits the extent to which conclusions could be drawn about the impact of some of the conditions on the equating results.

A second limitation of the resampling methodology was that a number of the effect size combinations were difficult to create for certain tests. Given that it was difficult to obtain a sample of examinees for some of the effect size combinations, obtaining multiple samples containing different examinees was not possible for those conditions. Therefore, bootstrap replications were used to create different samples of examinees rather than drawing new samples of examinees from the original full sample of examinees. Further, examinees were sampled based on demographic variables in an effort to eliminate correlated error that could result from sampling examinees based on their common-item scores. However, because examinees were sampled from demographic groups until the target common-item effect size was obtained, it is possible that correlated errors were still introduced into the sampling process.

Criterion equating relationships. In this study, the criterion equating relationship differed for each equating method. Because there were three criterion equating relationships, comparing bias across equating methods may not be advisable. Although the criterion equating relationships were similar across equating methods, differences existed. Therefore, although the IRT equating methods may have appeared to result in less bias relative to the traditional equating methods, this conclusion may have been due, in part, to the use of the different criterion equating relationships.

Future Research

A variety of issues concerning equating mixed-format tests were considered in this study. Although a number of limitations with the current research were discussed in the previous section, the diverse array of topics provides a rich starting point for future research. A number of the limitations in this study can likely be addressed only through a simulation study. A primary limitation of this study was use of the resampling methodology. The complex interactions among effect sizes for the various item formats illustrates the need to hold certain scores and effect sizes constant while manipulating others. A future simulation study should investigate alternative ways of calculating relative difficulty of MC and CR items. Specifically, examinees could be simulated with different proficiency levels across the two latent dimensions represented by MC and CR items. A simulation study could also incorporate various levels of MC and CR correlations for a given subject area. However, a major limitation of simulation studies is that it can be difficult to design them to realistically represent operational test situations. Therefore, careful attention will need to be given to the simulation study design.

Another area for future research is replicating this study on other tests or other subject areas. The authors have conducted similar research on tests of English Language and Chemistry with similar results (Hagge, 2010). However, characteristics of the tests limited the extent to which certain effect sizes could be obtained or representative sets of mixed-format common items could be created. Therefore, additional research is needed in order to generalize the results obtained in this study to other subject areas.

Conclusion

This study made use of a pseudo-test design to investigate issues associated with equating mixed format tests. The results of this study suggest that the test, examinee, and common-item characteristics investigated in this study impact equating results. Large differences in 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 equating method or inclusion of CR items in the common-item set. Specifically, inclusion of CR 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 criterion equating relationships in this study. Future research is needed in order to determine the specific conditions for which these

findings can be expected to hold and to develop guidelines that can be generalized to operational testing situations.

References

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

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, The 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 study, 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.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11-26.

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.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education, 14*, 31-57.

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

Hagge, S. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups.* Unpublished doctoral study. University of Iowa.

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

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 [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

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

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

Kolen, M. J. (2004). POLYEQUATE [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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

Lee, W., Hagge, S. L., He, Y., Kolen, M. J., & Wang, W. (2010, April). *Equating mixed-format tests using dichotomous anchor items*. Paper presented in the structured poster session Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple-Choice and Constructed-Response Items at the Annual Meeting of the American Educational Research Association, Denver, CO.

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

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

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.

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: John Wiley & Sons.

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.

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

Tan, X., Kim, S., Paek, I., & Xiang, B. (2009). *An alternative to the trend scoring method for adjusting scoring shifts in mixed-format tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://www.etsliteracy.org/Media/Conferences_and_Events/ AERA_2009_pdfs/AERA_NCME_2009_Tan.pdf

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

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

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

Walker, M. E., & Kim, S. (2009). *Linking mixed-format tests using multiple choice anchors.* Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://www.ets.org/Media/ Conferences_and_Events/AERA_2009_pdfs/AERA_NCME_2009_Walker.pdf

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). *Robustness in using multiple-choice items as an external anchor for constructed-response test equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Table 1

Descriptive Statistics for Spanish Language 2004

Item Format	Descriptive Statistics	Spanish Language 2004
	N (*Before Imputation*)	20,000
	N (*After Imputation*)	19,010
MC	Mean	60.561
	SD	15.148
	Skewness	-.479
	Kurtosis	2.625
	α	.933
CR	Mean	37.308
	SD	10.547
	Skewness	-.669
	Kurtosis	2.775
	α	.860
CI	Mean	16.927
	SD	4.919
	Skewness	-.332
	Kurtosis	2.445
	CO Corr.	.838
CO	Mean	97.869
	SD	24.470
	Skewness	-.589
	Kurtosis	2.751
	α	.947

Table 2

Composition of Pseudo-Test Forms

Form	Item Type	Number of MC (Points)	Number of CR (Points)	Difficulty		Discrimination	
				Mean	SD	Mean	SD
Old	Total Test	68 (68)	20 (40)	.648	.177	.376	.124
	NCR	33 (33)	0 (0)	.649	.151	.358	.097
	FCR	21 (21)	6 (12)	.650	.165	.364	.113
	Full CI	46 (46)	13 (22)	.631	.169	.370	.118
New	Total Test	68 (68)	20 (40)	.641	.167	.382	.136
	NCR	33 (33)	0 (0)	.649	.151	.350	.107
	FCR	21 (21)	6 (12)	.650	.165	.355	.129
	Full CI	46 (46)	13 (22)	.631	.169	.360	.135

Table 3

Summary of Factors of Investigation

CI ES	MC-CR ES	CI Format	Equating Methods
.00	.00	NCR, FCR	FE, CE, TS, OS
.00	.25	NCR, FCR	FE, CE, TS, OS
.20	.00	NCR, FCR	FE, CE, TS, OS
.20	.25	NCR, FCR	FE, CE, TS, OS
.40	.00	NCR, FCR	FE, CE, TS, OS
.40	.25	NCR, FCR	FE, CE, TS, OS
.60	.00	NCR, FCR	FE, CE, TS, OS
.60	.25	NCR, FCR	FE, CE, TS, OS

Table 4

Observed and Disattenuated MC and CR Correlations for Spanish Language

Form	Correlation	Single Group	MC-CR .00				MC-CR .25			
			CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
Original 2004	MC and CR	.808								
	Disattenuated MC and CR	.902								
Pseudo Old	MC and CR	.760	.781	.779	.786	.730	.776	.724	.746	.778
	Disattenuated MC and CR	.880	.893	.893	.896	.864	.889	.849	.865	.898
Pseudo New	MC and CR	.757	.747	.762	.791	.752	.740	.760	.766	.804
	Disattenuated MC and CR	.875	.860	.874	.899	.869	.866	.877	.884	.914

Table 5

Descriptive Statistics for the Old Spanish Language Pseudo-Test Form

Score	Statistic	Single Group	MC-CR .00				MC-CR .25			
			CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
	N	19,010	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	45.029	42.979	45.534	44.700	47.225	40.706	44.251	44.928	47.185
	SD	11.286	11.737	10.925	11.139	11.180	11.701	11.900	11.961	10.430
	Skew	-.401	-.317	-.383	-.424	-.556	-.217	-.364	-.388	-.494
	Kurt	2.624	2.431	2.654	2.611	2.863	2.328	2.387	2.511	2.816
	α	.909	.912	.904	.905	.912	.909	.918	.920	.898
CR	Mean	25.438	24.304	25.980	24.792	26.522	22.616	25.182	25.561	27.199
	SD	7.030	7.356	6.666	6.959	6.979	7.497	7.291	7.398	6.275
	Skew	-.614	-.502	-.701	-.422	-.733	-.373	-.693	-.658	-.904
	Kurt	2.897	2.634	3.239	2.548	3.140	2.482	2.992	2.904	3.886
	α	.821	.833	.803	.822	.823	.841	.830	.837	.795
CO	Mean	70.467	67.283	71.514	69.492	73.746	63.322	69.433	70.489	74.384
	SD	17.245	18.050	16.407	16.975	17.179	18.171	18.165	18.357	15.611
	Skew	-.517	-.400	-.559	-.447	-.660	-.318	-.517	-.522	-.691
	Kurt	2.809	2.545	3.046	2.662	3.059	2.376	2.686	2.722	3.264
	α	.929	.933	.923	.927	.931	.933	.935	.938	.920
NCR CI	Mean	21.423	20.439	21.641	21.327	22.524	19.348	21.041	21.407	22.365
	SD	5.855	6.040	5.763	5.812	5.789	6.003	6.111	6.087	5.439
	Skew	-.321	-.235	-.345	-.350	-.457	-.139	-.291	-.299	-.385
	Kurt	2.485	2.361	2.523	2.501	2.647	2.319	2.352	2.385	2.534
	CO Corr.	.913	.920	.905	.911	.919	.918	.920	.920	.902
FCR CI	Mean	22.139	21.111	22.459	21.829	23.182	19.817	21.864	22.147	23.244
	SD	5.705	5.931	5.489	5.629	5.609	5.924	5.982	6.059	5.170
	Skew	-.490	-.401	-.553	-.432	-.587	-.290	-.507	-.488	-.602
	Kurt	2.752	2.539	2.968	2.703	2.928	2.457	2.681	2.597	3.045
	CO Corr.	.934	.939	.925	.932	.935	.937	.940	.942	.920
Full CI	Mean	44.194	42.171	44.891	43.777	46.440	39.779	43.491	44.143	46.574
	SD	11.359	11.783	10.892	11.251	11.309	11.930	11.930	12.018	10.349
	Skew	-.455	-.340	-.520	-.416	-.600	-.286	-.460	-.443	-.588
	Kurt	2.700	2.467	2.930	2.607	2.925	2.344	2.608	2.602	3.010

Table 6

Descriptive Statistics for the New Spanish Language Pseudo-Test Form

Score	Statistic	Single Group	MC-CR .00				MC-CR .25			
			CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
	N	19,010	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	44.486	42.889	43.079	39.548	38.998	39.724	40.777	39.117	39.965
	SD	11.795	11.560	12.013	12.043	12.659	11.554	11.962	12.306	11.724
	Skew	-.398	-.320	-.297	-.083	-.047	-.102	-.177	-.083	-.133
	Kurt	2.541	2.497	2.405	2.269	2.157	2.376	2.409	2.255	2.384
	α	.914	.908	.916	.911	.920	.903	.911	.915	.906
CR	Mean	27.110	26.126	26.415	23.321	23.128	22.059	23.039	21.874	22.666
	SD	7.930	7.713	8.095	8.269	8.813	7.508	7.898	7.953	7.764
	Skew	-.732	-.724	-.663	-.180	-.169	-.063	-.118	-.019	-.093
	Kurt	2.662	2.787	2.539	2.113	2.002	2.481	2.299	2.266	2.296
	α	.818	.804	.820	.823	.841	.835	.833	.846	.828
CO	Mean	71.596	69.015	69.494	62.869	62.126	61.783	63.816	60.991	62.631
	SD	18.536	18.029	18.910	19.130	20.429	17.871	18.694	19.223	18.294
	Skew	-.575	-.529	-.475	-.149	-.118	-.095	-.164	-.063	-.139
	Kurt	2.677	2.661	2.488	2.229	2.103	2.447	2.403	2.277	2.383
	α	.930	.924	.931	.929	.937	.928	.932	.937	.928
NCR CI	Mean	21.423	20.572	20.826	19.293	19.009	19.637	20.091	19.279	19.649
	SD	5.855	5.733	5.946	5.933	6.131	5.872	6.004	6.258	5.924
	Skew	-.321	-.206	-.244	-.063	-.038	-.124	-.159	-.089	-.137
	Kurt	2.485	2.426	2.375	2.384	2.265	2.321	2.365	2.236	2.384
	CO Corr.	.891	.882	.892	.896	.908	.904	.907	.910	.895
FCR CI	Mean	22.139	21.173	21.577	19.860	19.566	19.934	20.481	19.631	19.997
	SD	5.705	5.648	5.843	5.899	6.198	5.716	5.890	6.128	5.812
	Skew	-.490	-.374	-.425	-.176	-.146	-.158	-.225	-.156	-.209
	Kurt	2.752	2.640	2.578	2.399	2.310	2.524	2.565	2.379	2.537
	CO Corr.	.926	.921	.931	.928	.935	.928	.930	.935	.924
Full CI	Mean	44.194	42.271	42.865	39.444	38.936	39.891	40.903	39.251	39.994
	SD	11.359	11.106	11.538	11.698	12.374	11.475	11.819	12.277	11.583
	Skew	-.455	-.372	-.382	-.140	-.119	-.153	-.213	-.120	-.170
	Kurt	2.700	2.617	2.562	2.351	2.232	2.460	2.441	2.268	2.436

Table 7

Effect Sizes for Spanish Language Pseudo-Test Forms

ES	Single Group	MC-CR .00				MC-CR .25			
		CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
N	19,010	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
CO	.063	.096	-.114	-.366	-.616	-.085	-.305	-.505	-.691
NCR	.000	.023	-.139	-.346	-.590	.049	-.157	-.345	-.478
FCR	.000	.011	-.156	-.342	-.612	.020	-.233	-.413	-.590
Full CI	.000	.009	-.181	-.378	-.633	.010	-.218	-.403	-.599
MC	-.047	-.008	-.214	-.444	-.689	-.084	-.291	-.479	-.651
CR	.223	.242	.059	-.193	-.427	-.074	-.282	-.480	-.642
MC-CR	-.270	-.250	-.273	-.252	-.262	-.010	-.009	.001	-.009

Table 8

Summary Statistics for Spanish Language Pseudo-Test Forms (NCR)

Statistic	Equating Method	MC-CR .00				MC-CR .25			
		CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
WARMSB	FE	1.033	1.146	1.738	2.542	3.366	3.986	4.417	5.420
	CE	.904	.929	1.219	1.576	3.398	3.837	4.027	4.681
	TS	.493	.519	.623	.416	2.758	2.681	2.847	3.035
	OS	.485	.463	.651	.431	2.805	2.719	2.866	3.051
WASE	FE	.521	.555	.554	.623	.542	.560	.597	.612
	CE	.615	.654	.646	.708	.603	.651	.689	.728
	TS	.467	.531	.521	.437	.423	.409	.453	.609
	OS	.450	.514	.505	.423	.410	.397	.443	.590
WARMSE	FE	1.157	1.274	1.824	2.617	3.409	4.025	4.457	5.454
	CE	1.093	1.136	1.379	1.728	3.451	3.892	4.085	4.738
	TS	.679	.743	.812	.604	2.790	2.712	2.883	3.096
	OS	.662	.691	.824	.604	2.835	2.748	2.900	3.107

Table 9

Summary Statistics for Spanish Language Pseudo-Test Forms (FCR)

Statistic	Equating Method	MC-CR .00				MC-CR .25			
		CI .00	CI .20	CI .40	CI .60	CI .00	CI .20	CI .40	CI .60
WARMSB	FE	1.128	.657	1.681	1.810	2.712	2.524	2.964	3.501
	CE	1.029	.466	1.277	1.123	2.655	2.177	2.520	2.714
	TS	.894	.454	.705	.432	3.023	2.747	3.079	3.075
	OS	.885	.399	.739	.440	3.067	2.792	3.130	3.125
WASE	FE	.482	.514	.511	.560	.503	.502	.534	.552
	CE	.560	.596	.588	.638	.553	.589	.608	.663
	TS	.374	.418	.403	.423	.411	.400	.413	.527
	OS	.359	.398	.390	.409	.398	.389	.402	.507
WARMSE	FE	1.227	.835	1.757	1.895	2.758	2.573	3.012	3.544
	CE	1.171	.757	1.405	1.291	2.712	2.255	2.593	2.794
	TS	.969	.617	.812	.605	3.051	2.776	3.107	3.120
	OS	.955	.563	.836	.600	3.092	2.819	3.155	3.166

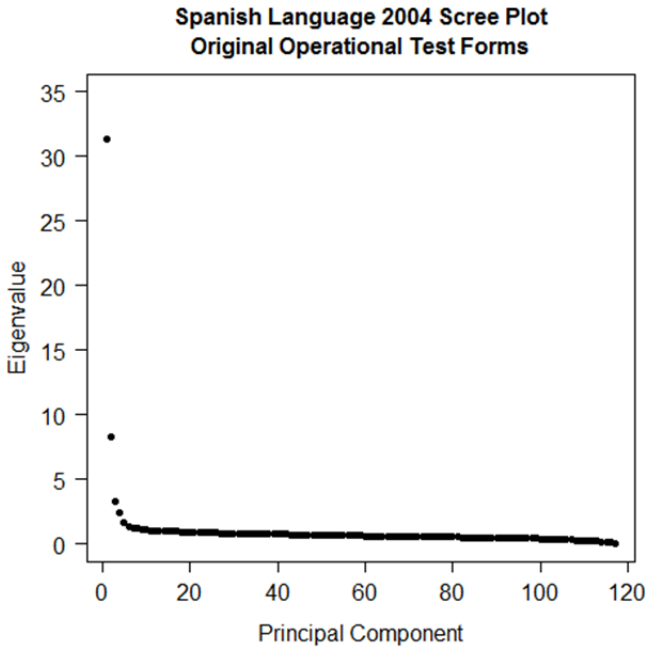

Figure 1. Spanish Language operational test form scree plot.

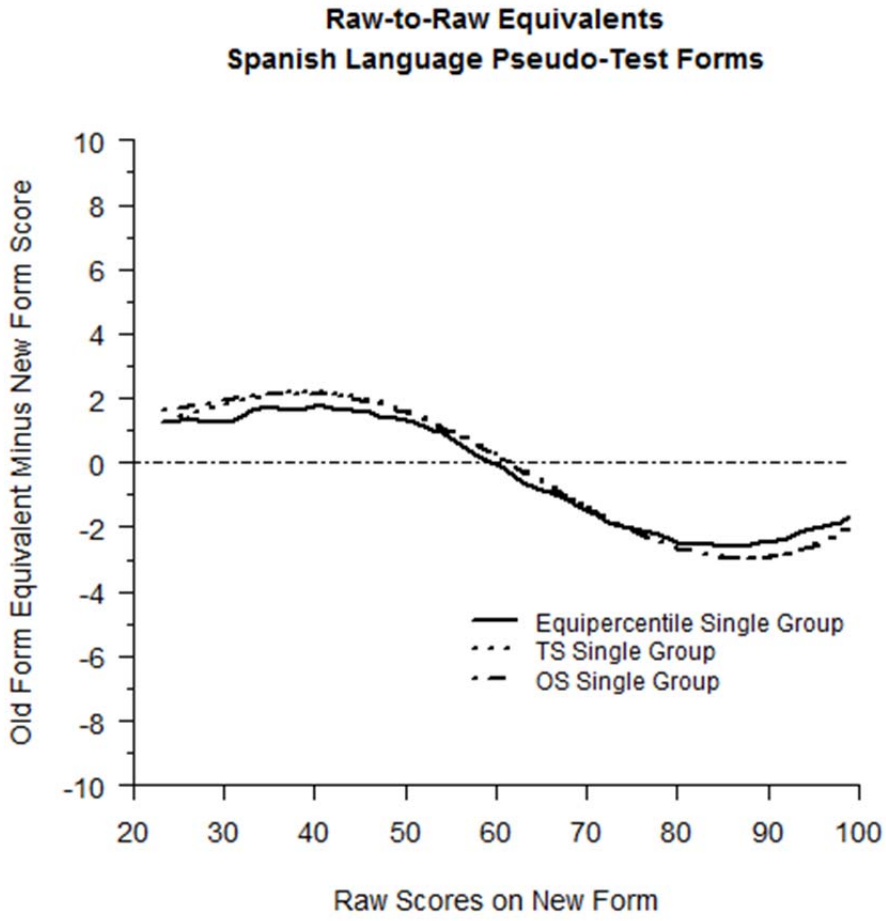

Figure 2. Criterion equating relationships for Spanish Language pseudo-test forms.

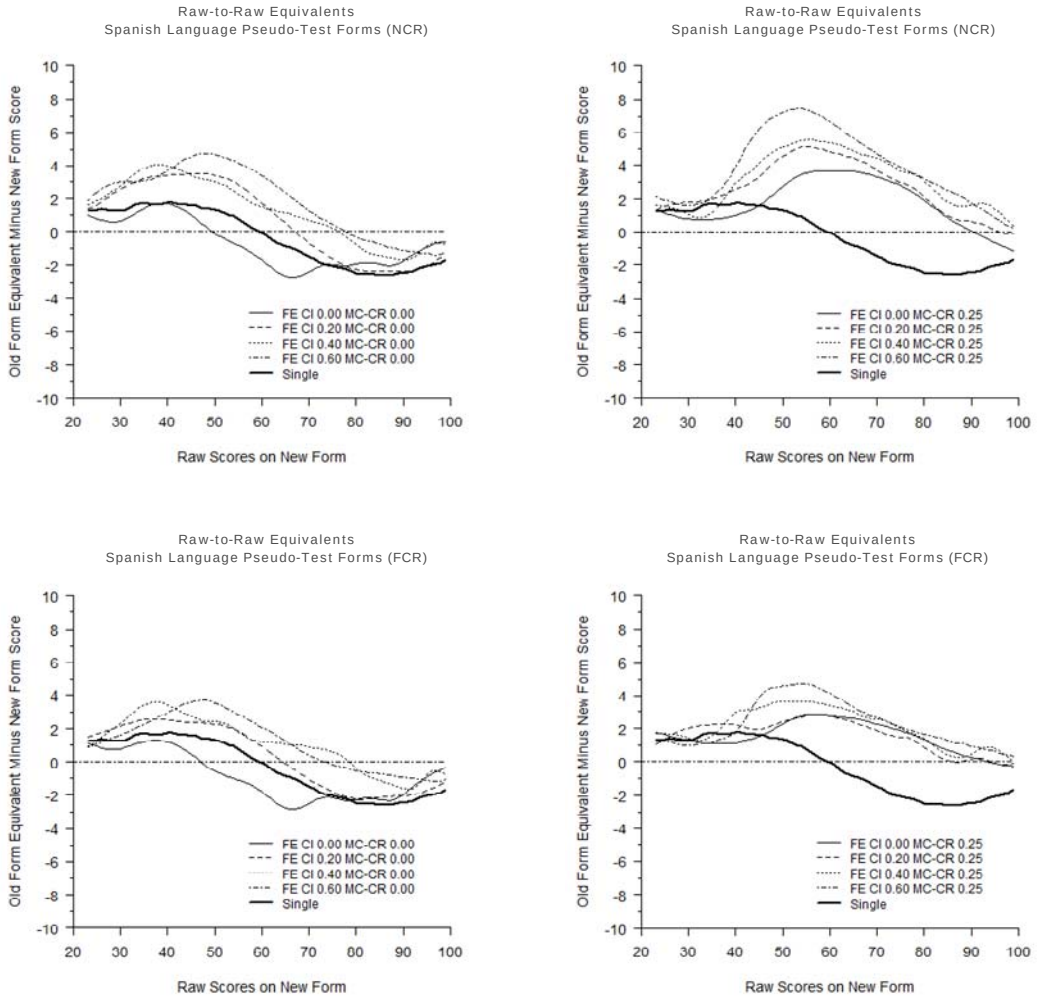

Figure 3. Spanish Language pseudo-test form comparison of CI ES for FE.

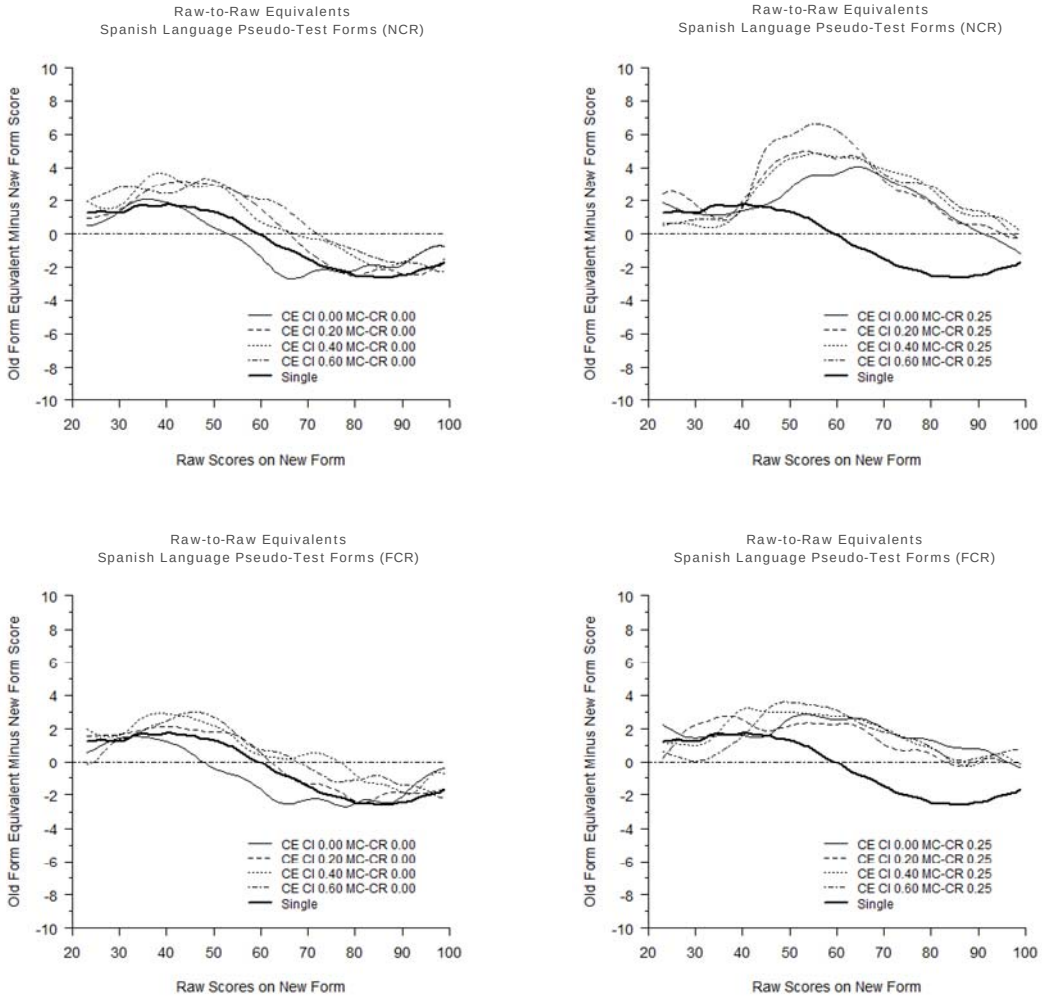

Figure 4. Spanish Language pseudo-test form comparison of CI ES for CE.

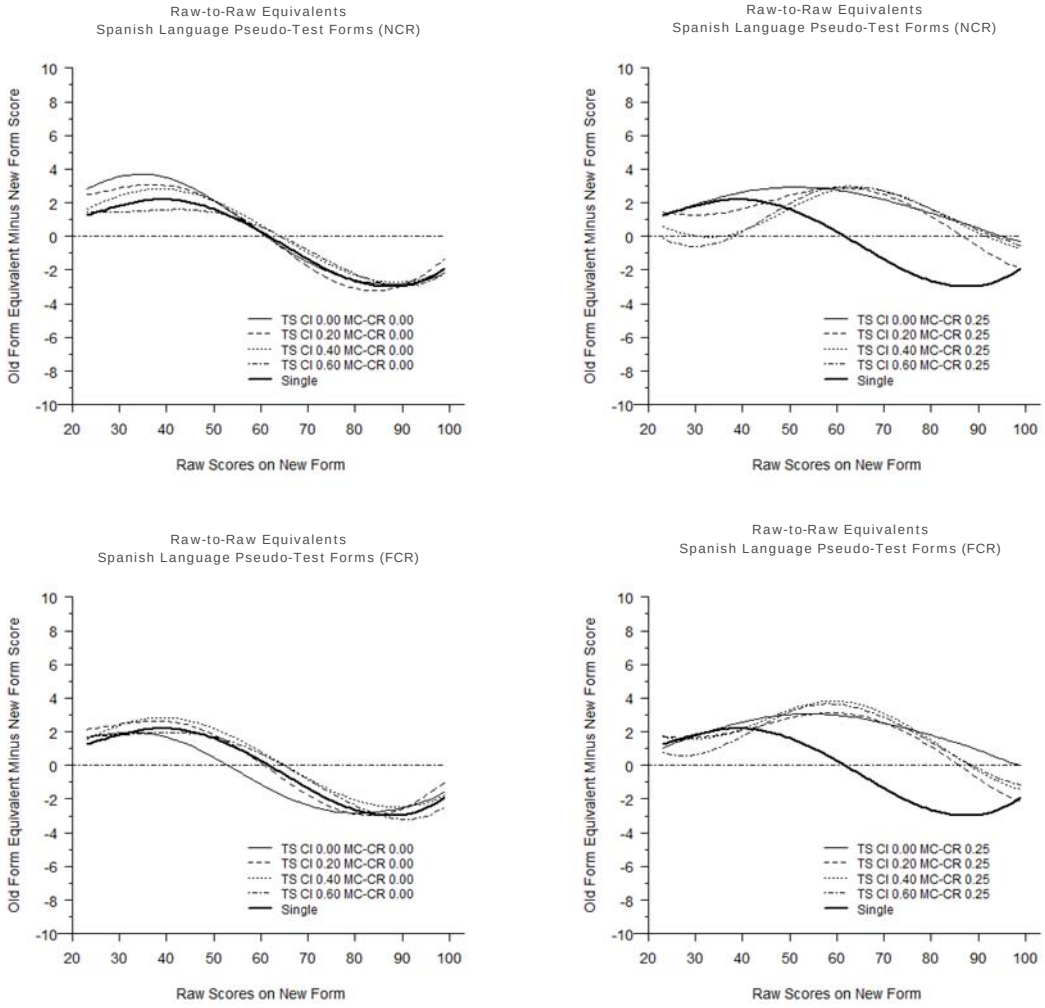

Figure 5. Spanish Language pseudo-test form comparison of CI ES for TS.

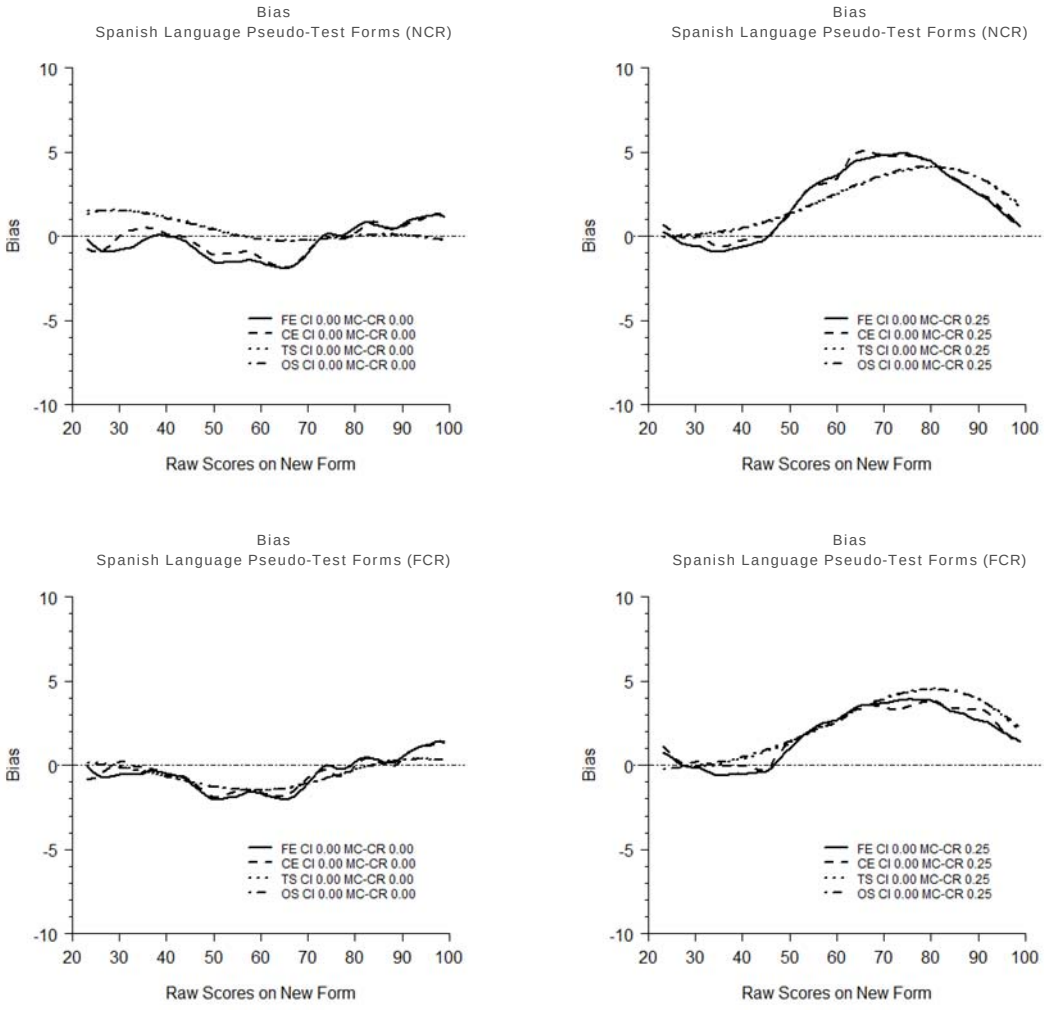

Figure 6. Spanish Language pseudo-test form conditional bias for CI .00.

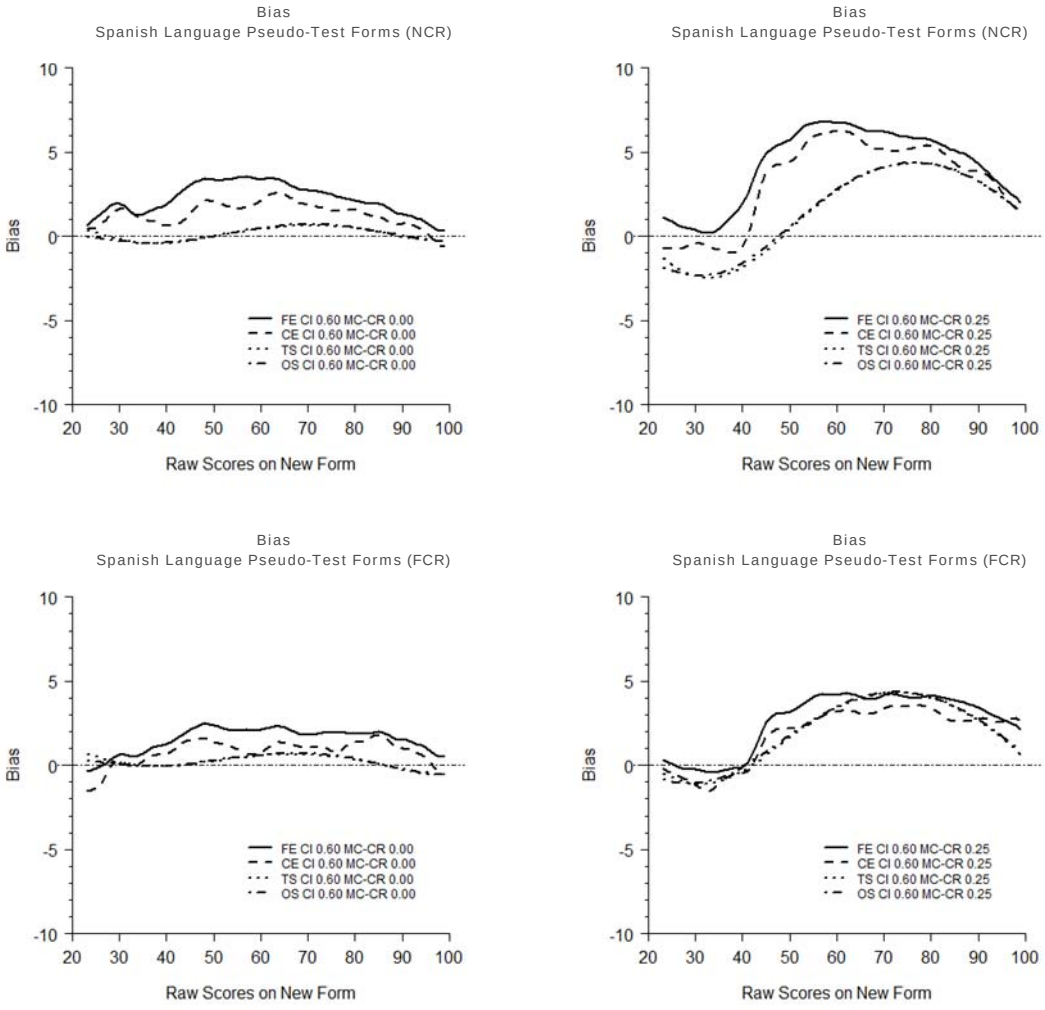

Figure 7. Spanish Language pseudo-test form conditional bias for CI .60.

Chapter 6: Evaluating Equating Accuracy and Assumptions for Groups that Differ in Performance

Sonya J. Powers and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

In order to compare examinee scores on multiple exam forms, it is essential that equating methods produce accurate results. Previous research has indicated that equating results may not be accurate when group differences are large. This study compared four curvilinear equating methods including frequency estimation, chained equipercentile, item response theory (IRT) true score, and IRT observed score equating. Using mixed-format test data, equating results were evaluated as group differences were systematically increased from no difference to a standardized mean difference (effect size) of .75. As group differences increased, equating results became increasingly biased and dissimilar across equating methods. Two hypotheses were evaluated to explain this finding: 1) Inaccurate equating results are caused by population dependent equating results; and 2) inaccurate equating results are caused by violations of equating assumptions. Findings from this study suggest that the cause of equating inaccuracies were violations of equating assumptions. Additionally, IRT and chained equipercentile equating methods appear to be less sensitive to group differences compared to the frequency estimation method. Results suggest that the size of group differences, the likelihood that equating assumptions are violated, and the equating error associated with a particular equating method, should be taken into consideration when choosing an equating method.

Evaluating Equating Accuracy and Assumptions for Groups that Differ in Performance

Different test forms that are built to the same content and statistical specifications are often administered to different groups of examinees for test security purposes and so that examinees can retake tests. Equating methods are used to adjust scores on test forms for differences in difficulty. When test forms are administered to examinee groups that differ in proficiency, equating procedures are used to disentangle group performance differences from differences in form difficulty.

The common item non-equivalent groups (CINEG) design provides a way to adjust for differences in form difficulty by imbedding a subset of items from a previous form into a new form. Common item scores are used to make adjustments for differences in form difficulty, taking into account differences in group performance. However, some research has indicated that the accuracy of CINEG equating results decreases as group differences increase (Kolen, 1990). According to Kolen and Brennan (2004, p. 286):

> ...mean differences between the two groups of approximately .1 or less standard deviation unit on the common items seem to cause few problems for any of the equating methods. Mean group differences of around .3 or more standard deviation unit can result in substantial differences among methods, and differences larger than .5 standard deviation unit can be especially troublesome.

There are at least two possible explanations for why large group differences might result in inaccurate equating results. One possibility is that the equating relationship between exam forms might differ for subgroups of examinees (i.e., equating relationships are population dependent). Another possibility is that the degree to which CINEG equating assumptions hold decreases as group differences increase. The purpose of this study was to investigate potential causes of equating inaccuracies observed when CINEG equating methods are used with old and new form groups that differ in performance.

Background

In this section, information is provided about the equating assumptions involved in single group (SG) and CINEG equating methods, a review of research comparing equating methods, and a review of research on population invariance.

Equating Assumptions

As with any statistical procedure, the accuracy of equating results depends on the extent to which the equating assumptions are met. Each equating method has its own set of assumptions that may or may not hold in a given situation. Additionally, various methods may be more robust than others to violations of assumptions. The following section describes the assumptions of traditional and item response theory (IRT) SG and CINEG equating methods.

In the SG design, one group of examinees takes both (or all) forms of an exam. The assumptions for the SG equipercentile equating method are (von Davier, Holland, & Thayer, 2004b):

1. There is a single population *P* of examinees that can take both forms.
2. A random sample from *P* is tested with both forms.
3. The order in which the forms are administered does not impact performance (i.e., there are no learning or fatigue effects).
4. Differences in difficulty between the forms can be adjusted by setting the distributions equal.

CINEG equipercentile equating methods considered in this study include chained and frequency estimation. Estimation of the quantities for the frequency estimation method is based on the assumption that the conditional distribution of total scores (*X* or *Y*) given the common item score (*V*) is the same in both populations. The assumptions for the chained method are that for a given population, the link from *X* to *V* is group invariant, and likewise that the link from *V* to *Y* is group invariant (von Davier et al., 2004b).

Although order effects are of concern with the SG design, the other traditional SG equating assumptions are less stringent than traditional CINEG equating assumptions, and may be less likely to be violated. Therefore SG equating results provide a good baseline for comparison with CINEG results.

Two SG IRT equating methods used with summed scoring are IRT true score equating and IRT observed score equating. The main assumption involved in unidimensional IRT is that responses to all items in the exam measure a single construct. With more than one form, scores on all forms must measure the same unidimensional construct. The unidimensionality assumption is considered stringent because of the multitude of factors that could plausibly affect student responses. An additional implicit assumption is that the IRT model chosen fits the data.

For IRT true score equating, there is also an implicit assumption that the relationship between true scores can be used with the observed scores (which are actually used). IRT true score and observed score equating methods can be used with the CINEG design with the additional assumption that the same construct is measured in the two groups of examinees and on both forms.

Comparison of Equating Methods

Although there are multiple equating methods, it seems desirable that all equating methods produce the same "true" equating relationship. However, equivalent results are unlikely to be obtained in practice because each equating method has its own underlying statistical assumptions, which hold to varying extents. Several studies have compared the equating results obtained using different CINEG equating methods.

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 have demonstrated comparable equating results across equating methods given optimal equating conditions (e.g., Cook, Dunbar, & Eignor, 1981; Cook, Eignor, & Taft, 1998; Marco, Petersen, & Stewart, 1979).

von Davier, Holland, and Thayer (2003, 2004a) provided a mathematical proof of the equivalence of the chained equipercentile and frequency estimation methods when old and new form groups are equivalent or given a perfect correlation between common item scores and total test scores. Although these two cases are extreme, they provide an indication that chained equipercentile and frequency estimation methods should provide reasonably similar results when group differences are small and when common item scores are highly correlated with total test scores.

Kolen (1981) compared the linear, equipercentile, IRT true score, and IRT observed score methods for 1-, 2-, and 3-parameter models. He found that IRT true and observed score methods with the 3-parameter model provided the most consistent results across two real-data replications, followed by the equipercentile method. Similarly, Han, Kolen, and Pohlmann (1997) compared the equating results of equipercentile, IRT observed, and IRT true score methods, using equating a test to itself as the criterion. The authors found that IRT true score equating was closest to the criterion, followed by IRT observed score, and then equipercentile. In

addition, Lord and Wingersky (1984) found that IRT true and observed score methods produced very similar results when equating a test to itself. Though neither method provided the identity equating relationship across the entire score scale, both methods were very close. Finally, comparing only the IRT equating methods under the CINEG design, Tsai, Hanson, Kolen, and Forsyth (2001) found that the standard error of IRT true score equating is slightly larger than the standard error of IRT observed score equating. These studies suggest that random equating error may be similar for IRT true and observed score equating methods, and that random IRT equating error maybe smaller than the random equating error produced by traditional equating methods.

Research suggests that the results of traditional chained and equipercentile equating methods under the CINEG design are not as similar as the results of IRT true and observed score equating methods. Harris and Kolen (1990) compared the chained and frequency estimation methods for groups that differed by approximately .1 and .3 standard deviations. The authors found that the two methods differed by more than the standard error (SE) of equating of the frequency estimation method. Differences between the two methods were greater when the standardized group difference was greater.

Sinharay and Holland (2007) noted that as group differences increased, so did the bias (but not the SE) in both frequency estimation and chained equipercentile methods. In their study, the chained method was always less biased than the frequency estimation method, a finding corroborated by Holland, Sinharay, von Davier, and Han (2008). However, the frequency estimation method may have slightly less random equating error compared to the chained equipercentile method (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008). Sinharay and Holland (2007) also found evidence that the equating error of the two methods is related to both the dimensionality of the data and the pattern of group differences across dimensions. Additionally, Ricker and von Davier (2007) found that the chained equipercentile method can have more bias than the frequency estimation method when the common item set is short relative to the total test length. The longer the relative length of the common item set, the less bias was found for both methods (Holland et al., 2008; Wang et al., 2008).

In a study by von Davier et al., (2003), the equating results produced using the chained equipercentile method were less population dependent than the results of the frequency estimation method. However, in a series of studies comparing linear, equipercentile, and IRT methods with different matched sampling designs, the sensitivity of the various methods to group

differences was called into question. Three studies found evidence that the frequency estimation method was less sensitive to group differences than chained equipercentile or IRT true score methods (Cook, Eignor, & Schmitt, 1988; Lawrence & Dorans, 1990; Schmitt, Cook, Dorans, & Eignor, 1990). Two other studies found opposite results (Cook, Eignor, & Schmitt, 1990; Eignor, Stocking, & Cook, 1990). Another study found that the sensitivity of equating methods to group differences differed across exams (Livingston, Dorans, & Wright, 1990). Finally, Wright and Dorans (1993) suggested that the type of matching methods used impacts the stability and accuracy of the equating results for different equating methods.

Population Invariance

If the equating relationship is the same for any subgroup of examinees from the same population, the equating results are considered population invariant (Kolen & Brennan, 2004). If an equating method produces a different equating relationship when applied separately to, for instance, males and females, using the population equating relationship and ignoring subgroup membership might put individuals from some of the subgroups at a disadvantage.

Several studies have looked at the degree to which equating results are population invariant in practice. Harris and Kolen (1986) compared the linear, equipercentile, and IRT true score methods with the random groups design and found that all methods provided reasonably population-invariant results. ETS conducted four separate studies of population invariance for Advanced Placement (AP) Exams (Dorans, 2003). The authors found that population invariance held for some subpopulations (e.g., geographical region), but not necessarily for others (e.g., racial groups). However, the degree to which subgroups differed in their equating/linking relationships depended on the specific exam (e.g., Calculus AB, Chemistry, etc.), the specific administration year, and to some extent, the equating method used. Finally, von Davier and Wilson (2006) conducted a population invariance study comparing Tucker linear, chained equipercentile, and IRT true score equating. The subgroups considered were males and females—groups that differed by approximately a quarter of a standard deviation. However, the within-subgroup year-to-year effect size (ES) was only .07. On average, differences between the subgroup equating relationships were not of practical significance.

Method

The source of the data, methods used to simulate group differences, and methods used to evaluate population invariance, differences in equating relationships, and equating assumptions, are described in this section.

Data Source and Cleaning Methods

Data from the 2008 form of AP English Language were included in this study. Multiple-choice (MC) items were operationally scored with a correction for guessing, also called formula scoring, to discourage examinees from randomly guessing. However, formula scoring tends to result in a substantial amount of missing data which is particularly problematic for IRT methods. Missing data were imputed using a two-way procedure described by Bernaards and Sijtsma (2000) and Sijtsma and van der Ark (2003). Only 31% of examinees responded to all items. To reduce the amount of missing data, examinees that responded to less than 80% of MC items were eliminated prior to the imputation process. Approximately 50,000 students were dropped in this step, a reduction of 17%. The percentage of missing scores was reduced from 11% to around 4%. The remaining 4% of missing values were imputed. 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 first four moments for the 55 item MC section of the test are provided in Table 1 for the entire sample, the sample with 80% response, and the imputed sample. The MC total score mean increased when examinees with low response rates were excluded and increased again when the data were imputed and number correct scoring was applied. The imputed data were less variable and more negatively skewed than the other two data sets.

Operationally, MC and FR sections were weighted by noninteger values to ensure that their contribution to the composite met test development requirements. However, use of noninteger weights results in noninteger scores for examinees. To avoid rounding noninteger scores, integer weights were used in this study. Integer weights of 2 for the MC section and 5 for the FR section were selected so that the sections would contribute to the composite score in the same ratios used operationally. The resulting scores ranged from 0 to 239.

Construction of Pseudo-Forms A and B

The 2008 AP English Language Exam form was divided into two pseudo-forms which are hereafter referred to as Form A and Form B. Although the examinees that took Form A and

Form B are the same, for the purposes of simulating the CINEG design, examinees that took Form A are referred to as the new form group, and the examinees that took Form B are referred to as the old form group. The "old" and "new" form groups are equivalent because the same examinees took both forms. Using the CINEG design in this artificial situation makes it possible to assess the statistical assumptions used with nonequivalent group equating methods.

MC items on the AP English Language Exam are all passage-based. All MC items tied to a single passage were considered part of a single testlet. MC testlets were assigned to Forms A and B randomly. It was sometimes necessary to include items on both Forms A and B so that the forms would contain equal numbers of items, and to keep items within a testlet together. FR items were assigned to both forms.

Common items included those used to link the operational 2008 form to previous AP English forms. Using the entire operational common item set (which included only MC items) with the shorter pseudo-forms resulted in a larger ratio (.67) of common items to total MC items than would typically be seen in practice (.34). However, all operational common items were used, because they should have been chosen to be as similar as possible to the MC items on the operational 2008 AP English Language Exam form in terms of content and statistical properties. Therefore, 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.

Sampling

A main objective of this study was to investigate the impact of group differences on equating results and assumptions. However, Forms A and B were taken by the same examinees. Therefore the common item effect size (ES) was exactly zero, because the common item scores were the same for the two forms. 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.

In order to investigate the impact of group differences on equating results, old and new form groups were sampled from the original group using each examinee's level of parental education. Parental education was selected for use in creating group differences because it had a

fairly low nonresponse rate, a large number of examinees within each category, and large ES differences between high and low categories in terms of common item scores. Parental education was defined as the highest level of education obtained by either parent. Levels were coded into the following categories (5% did not respond):

1. Any education through high school diploma or equivalent (11%)
2. Trade school, some college, or associates degree (19%)
3. Bachelors degree or some graduate/professional school (35%)
4. Graduate/Professional degree (30%)

Four old and new form groups were constructed by sampling examinees from the four levels of parental education. ESs considered included 0, .25, .5, and .75. A zero ES condition was created by splitting the full sample of examinees into two mutually exclusive stratified (by parental education) random samples. One sample was assigned to Form A, the other to Form B.

To obtain ESs greater than zero, the ratios of examinees within parental education categories were manipulated using a combination of brute force and Proc SurveySelect iterations. The process continued, modifying as necessary the number of examinees sampled from each level of parental education, until the resulting ES was again within .01 of the target ES. The overall sample size for Forms A and B was kept the same, and an attempt was made to keep sample sizes as large as possible across the four levels.

To obtain group differences, more examinees were sampled from the lower categories of parental education and fewer examinees were sampled from the higher categories of parental education for Form A. For Form B, more examinees were sampled from the higher categories of parental education and fewer examinees were sampled from the lower categories. Because of the relationship between examinee performance and parental education, this sampling technique resulted in higher performing old form groups.

The largest samples were obtained for the zero ES, and the sample size decreased as the ES increased. In order to keep the equating results comparable in terms of sampling error, a random sample of 1,500 was taken from each of the initial samples. Target and observed ESs and the percentage of examinees within each parental education category are listed in Table 2. The observed ESs based on the samples of 1,500 (Obs ES) were fairly close to the target ESs.

Equating Designs and Methods

The two equating designs considered in this study were the SG design and the CINEG design. The equating methods used with the SG design were the equipercentile method with cubic spline postsmoothing and the IRT true and observed score methods. CINEG equating methods included frequency estimation and chained equipercentile with cubic spline postsmoothing, and IRT true and observed score methods. Cubic spline postsmoothing was used in this study because it has been found to improve equating results and decrease the standard error (SE, Kolen, 1984; Hanson, Zeng, Colton, 1994). Smoothing parameters were chosen based on a visual inspection of the smoothed and unsmoothed equivalents. A smoothing value of S=.1 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 .5 were used for both new and old form groups.

MULTILOG (Thissen, Chen, & Bock, 2003) was used for IRT estimation for the three-parameter logistic (3PL) model with the MC items and for the graded response model (GRM; Samejima, 1972) with the FR items. For IRT observed score equating, normal quadrature points and weights were used to approximate the ability distribution for the old form group. Equating was conducted using *Equating Recipes*, a set of open source C functions (Brennan, Wang, Kim, & Seol, 2009). Simultaneous calibration was used for the SG design. For the CINEG design the Haebara (1980) scale transformation method was used to place parameter estimates on the same scale before equating.

SEs were estimated using a bootstrap procedure with 1,000 replications. SEs for traditional equating methods were calculated using the same smoothing value for each replication. IRT SEs are not estimated by *Equating Recipes*, and therefore they were not included in this study.

Population Invariance

Population invariance studies have typically selected one categorical examinee characteristic that is related to exam performance (e.g., gender) and equated forms separately for each group (e.g., males and females). Groups included in population invariance studies may differ substantially in performance from one another. However, within a particular group, the performance differences between administrations may be very small. For example, men and women may differ by a third of a standard deviation or more in terms of their scores on a given

exam. However, the performance of females from one administration to the next may be very similar. Therefore, the equatings conducted in population invariance studies typically involve very similar (or the same for the SG design) new and old form groups. However, the equating relationships compared are based on groups that differ substantially.

In order to investigate the degree of population dependence in the equating relationships for subgroups with different levels of parental education, equating was conducted separately for each level of parental education. The total group equating relationship was used as the criterion equating relationship, where the total group represented a stratified (by parental education level) sample of 1,500 examinees.

Because the purpose of the population invariance analysis was to determine the extent to which equating results were population dependent, only the SG design was used so as not to confound population invariance with violations of the more stringent CINEG equating assumptions. The direct equipercentile, IRT true score, and IRT observed score equating methods were used to estimate equating relationships for each of the parental education levels and the total group.

The SEs for the direct equipercentile equating method were calculated for the total group so that comparisons of equating relationships could be judged relative to random equating error. Presumably, if an equating relationship was within two SEs of the total group equating relationship, then the differences may have been caused by random error, not population dependence. Using equipercentile SEs to compare IRT equating results may be liberal (Liu & Kolen, 2010). However, using SEs rather than the standard error of the equating difference (SEED; von Davier et al., 2004b), which takes into account more sources of variability than the SE, may have led to overly conservative comparisons. However, the SEED requires a more complicated bootstrapping procedure that is not a standard function in *Equating Recipes*. Therefore, the unsmoothed equipercentile SEs for the total group were used to compare subgroup and total group equating relationships.

Root expected mean squared difference (REMSD) statistics, which are defined later in this section (Dorans & Holland, 2000; von Davier et al., 2004a), and classification consistency of AP grades (1 to 5) were also calculated to compare total group and subgroup equating relationships. Classification consistency (CC) was calculated in terms of the grade an examinee

would get for the total group equating relationship compared to the grade an examinee would get based on the equating relationship obtained for a given level of parental education.

$$CC = \frac{n_{11} + n_{22} + n_{33} + n_{44} + n_{55}}{N}, \qquad (2)$$

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 on the composite score scale for the AP grades were determined for Form B (the old form), by selecting the integer composite score plus .5 that corresponded most closely to the cumulative frequency of examinees at or below the cut score on the operational form. In other words, if the cut score for the operational AP English Language Exam had 10 percent of examinees at or below an AP grade of 2, and if 10 percent of examinees had a composite score less than or equal to 60 on Form B, then 60.5 was selected as the minimum composite cut score required to be classified into the AP grade 2 category. This procedure resulted in approximately equal numbers of examinees receiving each AP grade under formula scoring and under imputed number correct scoring.

Because the AP grade an examinee receives is what matters in terms of college credit or advanced placement in the AP examination program, classification consistency provides an indication of the practical significance of population dependent equating results. Low classification consistency for different levels of parental education, provides strong evidence that population invariance does not hold.

Equating with Group Differences

Whereas the old and new form groups in population invariance studies are often very similar, another possible comparison involves equating relationships for dissimilar old and new form groups. However, when equating using nonequivalent groups, differences in equating relationships can be caused not only by population dependence, but also by violations of the statistical assumptions involved with the CINEG equating methods.

New and old form groups were created to systematically differ by an ES of approximately 0, .25, .5 and .75. 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 criterion equating relationship were also used to evaluate whether comparison group

equating relationships were within sampling error of the criterion relationship. However, SEs were only calculated for the frequency estimation and chained equipercentile methods.

REMSD was used to provide an overall indication of how equating relationships departed from the criterion relationship as ES increased.

$$REMSD = \frac{\sqrt{\sum_{\min(a)}^{\max(a)} v_{ac}[eq_{Bc}(a) - eq_{B0}(a)]^2}}{\sigma_0(B)} \quad (3)$$

Here v_{ac} is the conditional proportion of Form A examinees at a particular Form A score for the comparison (c) equating relationship; $eq_{B0}(a)$ the is Form B equivalent for the ES of zero condition; $eq_{Bc}(a)$ is the Form B equivalent for a comparison 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. The equating method used with the criterion group (ES = 0) was also used with the comparison group (ES = .25, .5, or .75).

Classification consistency was calculated to provide an indication of the practical significance of equating differences. It 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).

Evaluating Equating Assumptions

Direct evaluation of the statistical assumptions of traditional equating methods used with the CINEG design can only be done when the responses of examinees have been observed for both forms. Examinee responses were observed for both forms in this study because a single operational form was divided into two pseudo-forms.

Frequency Estimation Method. Direct evaluation of frequency estimation method assumptions involves comparisons of v +1 pairs of conditional distributions, where v is the number of common items. The comparison is between the conditional distributions of Form A scores given common item scores in the two groups [$f_1(A|V)$ vs. $f_2(A|V)$], and between the conditional distributions of Form B scores given common item scores in the two groups [$g_1(B|V)$ vs. $g_2(B|V)$]. The weighted average of v +1 maximum differences in the cumulative frequency distributions, across all v +1 pairs of conditional distributions, was calculated for each equating relationship for Form A and Form B.

Chained Equipercentile Method. Unsmoothed equipercentile equivalents were calculated for the link from *A* to *V* [$e_V(A)$] and from *V* to *B* [$e_B(V)$] in the old (2) and new (1) form groups to evaluate the chained equipercentile method assumptions. Differences in the linking relationships for the two groups were quantified using REMSD statistics.

IRT True and Observed Score Methods. IRT true and observed score equating methods also have statistical assumptions which can be evaluated, albeit less directly. To investigate whether or not IRT model assumptions 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

This section describes the outcomes of the population invariance analyses, comparisons of equating relationships, and evaluations of equating assumptions.

Population Invariance

Equating results for the three SG equating methods and four parental education subgroups (L1-L4) are displayed in Figure 1**.** The top plot provides results for the IRT true score equating method. Symmetric bold lines represent plus and minus two SEs for the unsmoothed equipercentile equating method, calculated using 1,000 bootstrap replications. Equating relationships for parental education level subgroups are indicated with lines with circles (L1), lines with triangles (L2), lines with squares (L3), and dashed lines (L4). Similar plots are included for the IRT observed score equating method and the equipercentile method. The closer the subgroup equating relationship is to a vertical axis value of zero, the closer it is to the total group (criterion) equating relationship for that equating method. Equating relationships are only plotted between the .5 and 99.5 percentiles because beyond these ranges there is very little data that can be used to estimate the equating relationships.

For the IRT methods, the subgroup results were within two (equipercentile) SEs of the criterion for the entire score scale. The postsmoothed equipercentile results were much less smooth than the IRT results. The equipercentile equating results for all subgroups were within two SEs of the criterion for the majority of the score scale. However, the equipercentile results

exceeded two SEs at several points. Increasing the smoothing parameter would have made the postsmoothed results less bumpy and may have kept the equipercentile equating results within the SE bands for more score points. However, increasing the smoothing parameter can also increase the bias in equating results.

Composite score means and standard deviations are provided in Table 3 for old form equivalents based on unsmoothed (UEq) and postsmoothed equipercentile (SEq), IRT true score, and IRT observed score equatings, for the total group and all four subgroups. Across equating methods the means increased as parental education level increased. Within each level of parental education the moments were very similar across equating methods.

REMSD values, used to quantify the difference between each criterion and comparison equating relationship, are provided for all three SG equating methods in Table 4. To facilitate interpretation of the REMSD values, often the standardized difference that matters (SDTM; Dorans, 2003) is used. SDTM is defined as a half of a reported score point divided by the old form standard deviation for the total group.

The equipercentile REMSD values were greater than the SDTM, even when smoothing was used. However, the IRT true and observed score REMSD values were less than the SDTM for parental education levels three and four. These findings suggest that IRT equating results may be less population dependent than equipercentile equating results.

To calculate classification consistency, several sets of cut scores were needed:

- *Operational AP Cut Scores.* These are the cut scores set by the College Board and used with the operational AP Exams (see “Operational” columns in Table 5).
- *Form B Cut Scores.* These were set so that the Form B percentages at each AP grade would be as similar as possible to the percentages that were operationally used for the total group (see “% Below” columns in Table 5).
- *Total Group Form A Cut Scores.* These were obtained using the total group equating relationship between Form A and Form B. Cut scores were found for each equating method (see “Total” column in Table 6).
- *Subgroup Form A Cut Scores.* These were obtained using the subgroup equating relationships between Form A and Form B. Cut scores were found for each equating method (see “Subgroup” columns in Table 6).

A comparison of the total group and subgroup Form A cut scores in Table 6 reveals that the cut scores were fairly consistent across subgroups and equating methods. Table 7 provides the classification consistency values where the total group Form A cut scores were considered the criterion values to which each subgroup Form A cut score was compared. Classification consistency was above 97% for all subgroups and equating methods. The IRT equating methods had higher classification consistency than the equipercentile equating method.

Equating with Group Differences

Equating results were compared for new form groups that differed incrementally from old form groups in terms of common item performance. Figure 2 provides a comparison of the equating relationships for different ESs for the four equating methods. The criterion equating relationship (ES=0) is represented by the vertical axis value of zero**.** The differences between the criterion and comparison equating relationships for each ES are indicated with lines with circles (ES=.25), lines with triangles (ES=.5), and lines with squares (ES=.75). Symmetric bold lines represent plus and minus two SEs for the chained equipercentile criterion (ES=0) equating relationship, calculated using 1,000 bootstrap replications. Although frequency estimation bootstrap SEs were also calculated, they were similar to (albeit slightly smaller than) the chained SEs. To simplify comparisons, only chained SEs were plotted.

A comparison of the plots in Figure 2 indicates that equating relationships for all ES values deviated from the criterion equating relationship by more than two SEs for part of the score range. The equating results for frequency estimation were the farthest from the criterion. The deviation of the comparison equating relationships from the criterion equating relationship was in the order expected: as the ES increased, the deviation from the criterion increased. This occurred for all four equating methods.

Figure 3 provides a comparison of the four equating methods at each ES. In Figure 3 the vertical axis value of 0 is the Form B equivalent minus the Form A score, not the criterion equating relationship as in the previous figures. Therefore, the SEs, provided only for the chained equipercentile equating relationship for reference, are not symmetrical about the horizontal line at the vertical axis value of 0. At ES=0, the traditional and IRT equating methods deviated at the lower end of the score scale somewhat, but were within plus or minus two chained equipercentile SEs for the entire score scale. The IRT observed and true score methods were very similar to one another for all ESs except at extreme scores. As ES increased, the

results for the traditional methods became less similar to one another and less similar to the IRT equating results.

The mean and standard deviation are provided for each equating method and ES combination in Table 8. Because the old form group was sampled to be higher performing than the new form group, the old form equivalent means generally decreased for all equating methods as the ES increased. At an ES of 0, all methods produced similar equated score moments. The two IRT methods had very similar moments at each ES, even .75. However, as the ES increased, the means became less similar for the traditional methods. The traditional moments also became less similar to the IRT moments as ES increased. The frequency estimation method means became increasingly higher than the IRT method means as ES increased. The means for the chained method tended to fall in between the frequency estimation and IRT means.

REMSD and SDTM values are provided for all ESs and equating methods in Table 9. The values for frequency estimation and chained equipercentile are based on the smoothed equivalents. The REMSD values increased as the ES increased, as expected. A comparison of Tables 4 and 9 indicates that the CINEG equating results for groups that differed in common item performance resulted in REMSD values that were larger than those obtained for the SG equating results. The larger REMSD values for the CINEG equating results may indicate that the equating assumptions for the CINEG equating methods have been violated to some extent.

Classification consistency values are provided in Table 10, where ES=0 is the criterion equating relationship, and each equating method is compared to itself. The highest classification consistency was found for the lowest comparison ES (.25) for all four equating methods. Classification consistency was particularly low for the frequency estimation method, and the values decreased as the ES increased. Classification consistency values were higher for the IRT equating methods than for the traditional equating methods.

In Table 11, classification consistency values are provided using IRT true score equating results as the criterion. Classification consistency in Table 11 does not indicate the accuracy of the equating methods, but rather the consistency of examinee classifications for the four equating methods. IRT methods provided very similar classifications. Classifications based on the traditional equating methods were less similar to the IRT true score classifications. Classification consistency decreased as group differences increased for the frequency estimation method. However, this relationship was not found consistently for the other equating methods.

Evaluating Equating Assumptions

Frequency Estimation Method. The weighted averages of the maximum differences in the cumulative frequency distributions for old and new form groups across ESs were very similar for scores on Form A given V and scores on Form B given V. Therefore the weighted averaged were averaged across forms. These values are provided in Table 12 (see FE column). The higher the values, the less well frequency estimation assumptions held. As ES increased, the weighted maximum differences increased. In other words, as old and new form groups became less similar, the degree to which frequency estimation assumptions held decreased. A comparison of the values in Table 12 with the frequency estimation REMSD values in Table 9 indicates that the degree to which the frequency estimation assumptions hold is related to the accuracy of the frequency estimation equating results.

Chained Equipercentile. REMSD values were very similar for the A to V relationship and the V to B relationship, so the REMSD values were averaged across and are provided in Table 12 (see column CE). Higher REMSD values indicate that the chained assumptions did not hold as well. REMSD values increased consistently as ES increased. A comparison of the REMSD values in Table 12 to the chained REMSD values in Table 9 indicates that the degree to which the chained assumptions held was directly related to the accuracy of the chained equating results.

IRT. Correlations and reliability estimates were fairly similar across ESs. The MC and FR reliabilities were moderate (.86 and .66 respectively), and the disattenuated correlation averaged .68. These results provided some evidence that the AP English Language Exam may not meet the unidimensionality assumption. There was no indication that the magnitude of group differences had any impact on the degree to which the unidimensionality assumption held.

Eigenvalues and the percentage of variance explained for the first three principal components are provided in Table 13 for Form A (Form B results are very similar). The first principal component accounted for less than 29% of the variance under all effect size conditions. Although a substantial proportion of the variance remained, the amount of variance explained by the second principal component was less than 5%, and the ratio of the second principal component to the third was relatively small. There did not appear to be a pattern between the ES and the magnitude of the eigenvalues or the percentage of variance explained.

Scree plots were provided for all principal components for the Form A ES=0 condition (see Figure 4). The scree plots for other ESs were nearly identical. The larger plot provides the eigenvalues for all principal components. The smaller plot, in the top right, provides the eigenvalues for all but the first principal component. There was a clear break in the eigenvalues of the large plot between the first and second principal components for Forms A and B. However, there also appeared to be a small break between the second and third principal components in the smaller plots. These results indicate that English Language may have one dominant dimension and one secondary dimension.

All MULTILOG runs converged well before reaching the maximum of 500 iterations. For the MC items, the item parameter estimates were all within the expected ranges. FR items had very low and high b-values for some categories, but the results appeared otherwise reasonable.

Item parameters were estimated for MC and FR items simultaneously and separately to determine the stability of parameter estimates. Scatter plots of item parameters estimated with and without scores from the other section are provided in Figure 5. Scatter plots were only included for Form A ES=0 because the results were similar across ESs and pseudo-forms. For the MC item parameter estimates, inclusion of FR items tended to increase the *c*-parameter estimates when compared to the MC-only item parameter estimates for Forms A and B. However, estimation of the MC *a*- and *b*-parameters appeared stable. The FR *a*-parameters were consistently higher when estimated without MC items. For each FR item, there were the number of categories minus one *b*-parameter estimates. These values were plotted with circles. The FR *b*-parameter estimates were more extreme when estimated with the MC items.

Discussion

This section provides a summary of results, as well as a discussion of the significance of the study, the limitations of the study, and areas for future research.

Summary of Results

SG equating methods were used to assess population invariance of equating relationships for parental education level subgroups. Comparisons of equating relationships, REMSD values, and classification consistency values indicated that the equating relationships for parental education subgroups showed little population dependence. In this study, IRT true and observed score equating methods consistently showed slightly less population dependence than the

equipercentile equating method. However, previous research (von Davier & Wilson, 2006; Dorans, 2003; Harris & Kolen, 1986) has not always found noticeable differences in the population invariance of IRT and traditional equating methods, so additional study is needed.

For the CINEG design, comparison of equating relationships, equated score moments, and classification consistency values indicated that IRT true and observed score equating results were very similar to one another even at an ES of .75. Lord and Wingersky (1984) also found that the two IRT equating methods provide very similar results. However, the IRT equating results became more biased and less similar to the traditional equating results as group differences increased. Likewise, equating results for postsmoothed frequency estimation and chained equipercentile became less similar and increasingly biased as ES increased. Similar findings have been reported by Harris and Kolen (1990) and Sinharay and Holland (2007). These differences cannot be attributed to population invariance, because, as previously mentioned, the population invariance analyses indicated minimal population dependence. Therefore, the inaccuracy of the CINEG equating relationships may have been caused by violations of equating assumptions. To test this hypothesis, traditional and IRT equating assumptions were evaluated.

As old and new form group performance differences increase, it seems likely that the groups may differ in more complex ways than just on total test scores. These group differences might result in groups for whom the common items perform differently in relationship to the total test score. Both traditional and IRT assumptions hinge on the common item-total test relationship remaining constant across groups. For the frequency estimation method, the assumptions involve the conditional distribution of total test scores given common item scores in both groups. For the chained equipercentile method, the assumptions involve the linking relationship between total test scores and common item scores in both groups. For IRT the assumption is that both forms and the common items measure the same unidimensional construct for all groups.

As group differences increased, the degree to which frequency estimation and chained equipercentile equating assumptions were violated increased, and equating bias increased. This finding helps to explain the general finding in the literature that group differences often lead to inaccurate and inconsistent equating results (e.g., Cook et al., 1981; Cook et al., 1998; Harris & Kolen, 1990; Kolen, 1990; Marco et al., 1979).

Although a variety of model fit and dimensionality analyses all indicated possible multidimensionality in the AP English Language Exam, no discernable relationship was found between the magnitude of group differences and the degree to which IRT assumptions held. However, IRT equating results were increasingly biased as group differences increased making it seem likely that group differences and IRT assumptions are related in some way not clearly illuminated using the analyses conducted in this study, especially given that IRT single group equating results proved to be more population invariant than the traditional equating results.

Significance of the Study

Research has shown that large group differences can cause inaccurate and inconsistent equating results (Kolen, 1990). However, none of the existing research has investigated the cause of equating inaccuracies when groups differed. This study used real data to investigate whether or not population dependence and/or violations of equating assumptions were responsible for the inaccurate or inconsistent equating results. Isolating the cause of equating inaccuracies when groups differ may help to identify methods for improving equating accuracy.

In many studies of population invariance, subgroup equating relationships were compared for CINEG equating methods (e.g., Dorans, 2003; von Davier & Wilson, 2006). However, the CINEG methods involve strong statistical assumptions. When those assumptions do not hold, equating relationships can differ even when equating relationships are not population dependent. This study disentangled the effects of population invariance and violations of statistical assumptions by conducting population invariance analyses using SG equating methods which have minimal assumptions. Moreover, previous population invariance research has focused primarily on the difference-that-matters (DTM) criterion for assessing population invariance (e.g., von Davier & Wilson, 2006; Dorans, 2003). In this study SEs were used as a way to distinguish significant subgroup equating differences from differences within sampling error of the total group equating relationship. Classification consistency was also calculated because AP grades are the scores that matter to examinees. With the subgroups considered in this study, the results indicate that the degree to which traditional equating assumptions were violated corresponded directly to the inaccuracy of equating results. There was only a relatively minor amount of population dependence of equating results for parental education subgroups despite the large performance differences between the groups.

In this study, ESs were manipulated to range from 0 to .75. Operational AP Exams have group differences of less than approximately .15 standard deviations from year to year. Even the highest ES reported for different SAT administration groups in Lawrence and Dorans (1990) was only approximately .40. However, the more extreme range of group differences presented in this study allows the impact of group differences on equating results to be seen more clearly. Research indicates that the impact of group differences on equating accuracy may be obscured by sampling error when group differences are smaller than an ES of .3 (Powers, 2010). This may explain why several studies carried out by ETS produced conflicting results when comparing the sensitivity of equating methods to group differences (Cook et al., 1988; Cook et al., 1990; Eignor et al., 1990; Lawrence & Dorans, 1990; Livingston et al., 1990; Schmitt et al., 1990; Wright & Dorans, 1993).

Four curvilinear equating methods were included in this study so that the sensitivity of equating methods to group differences could be compared for more extreme ESs. Results suggest that the IRT and chained equating results may be less sensitive to group differences than the frequency estimation method.

An additional finding in this study was that the SE for the frequency estimation method was smaller than the SE for the chained equipercentile method. This finding has been noted in previous research (Sinharay & Holland, 2007; Wang et al., 2008). Although SEs for IRT equating methods were not estimated in this study, research by Liu and Kolen (2010) suggests that the IRT equating methods may have smaller SEs than traditional curvilinear equating methods.

Based on these findings, it appears that IRT true and observed score equating methods may have less random and systematic sources of equating error compared to the chained equipercentile and frequency estimation equating methods. The choice between the chained and frequency estimation methods depends on the size of group differences. If group differences are small, the frequency estimation method may have little bias, and less random error compared to the chained equipercentile method. Additionally, Ricker and von Davier (2007) found that the frequency estimation method may have less bias than the chained equipercentile method when the common item set is relatively short compared to the total test length. However, given sufficient numbers of representative common items, the chained method appears less sensitive to large group differences. Taken together, these results suggest that selection of an equating

method should take into consideration the size of group differences, the likelihood that equating assumptions have been violated, and the random error associated with a particular equating method.

Limitations of the Study

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. Inclusion of additional samples and additional tests are needed so that the results found in this study can be replicated.

Though the purpose of this study was to investigate the relationship of group differences and equating accuracy, all of the data came from one AP English Language Exam form. In addition, a number of modifications were made to the scores and groups used in this study, which may limit the generalizability of the results. Also, the initial sampling to obtain "old" and "new" form groups that differ by target ESs created nonequivalent groups that would be unlikely to have occurred in practice. Specifically, parental education was chosen to be the sampling variable because of its close relationship to examinee performance. The old form group was sampled from the original group by including more examinees from higher levels of parental education, and the new form group was sampled from the original group by including more examinees from the lower levels of parental education. However, in practice, different administration groups are unlikely to differ in terms of parental education as substantially as they did in this study. Parental education was chosen so that group differences could easily be manipulated. However, this study does not investigate the impact of group differences on equating accuracy when groups differ on a variable that is less related to exam performance.

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 are known to produce less biased equating results (Kolen, 1990). A change in the composition of the common item set 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 (Kim, Walker, & McHale, 2008).

Issues of composite-to-common item ratios were also not considered. Linear equating methods, multi-dimensional IRT methods, and IRT testlet models were not investigated. 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.

Future Research

Because of the limited scope of this study, there are several areas for future research. Several more equating models could be considered including traditional linear equating methods, multiple smoothing values, multidimensional IRT methods, and IRT methods for testlets.

Including additional exams with large ESs would improve the generalizability of the findings presented in this study, especially if the exams were from different testing programs, different content areas, and taken by different types of examinees. With a greater variety of exam and examinee types, it will be possible to determine whether or not the relationship of group differences and equating accuracy is predictable across exams.

Inclusion of more samples for each ES level would allow for the estimation of sampling error as well as the bias in estimation. In future studies, equating relationships could be compared using the more appropriate SEED statistic (von Davier et al., 2004b) instead of SEs. SEED incorporates sampling error in both the criterion and comparison equating relationships and can be used to determine which comparison equating relationships diverge from the criterion relationship by more than sampling error. Although chained equipercentile SEs were used to compare equating relationships in this study, the appropriate SEs or SEEDs should be used to make comparisons for each equating method, including IRT methods.

Exams that are scored number correct operationally could be used to avoid the need for imputation. Also, single-format exams as well as mixed-format exams could be investigated, with more attention given to the number of common items, and the representativeness of the common item set in terms of both content and item format.

Conclusions

In order to compare examinee scores on multiple exam forms, it is essential that equating methods produce accurate results. The results found here and in other research (Kolen, 1990)

indicate that equating results may not be accurate when group differences are large. Findings from this study suggest that the cause of equating inaccuracies may be violations of equating assumptions. IRT and chained equipercentile equating methods appear to be less sensitive to group differences compared to the frequency estimation method. However, when group differences are small, equating bias may be much less of a concern than random equating error. As in other studies (Sinharay & Holland, 2007; Wang et al., 2008), the SE for the frequency estimation method was found to be smaller than the SE for the chained equipercentile method. Some research indicates that IRT equating methods have smaller SEs than traditional equating methods (Liu & Kolen, 2010). These results suggest that the size of group differences, the likelihood that equating assumptions are violated, and the equating error associated with a particular equating method, should be taken into consideration when choosing an equating method.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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.

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.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus post-stratification equating methods. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program Examinations* (ETS Research Report RR-03-27, pp. 19-36). Princeton, NJ: Educational Testing Service.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating.* New York: Springer-Verlag.

von Davier, A. A., & Wilson, C. (2006). Population invariance of IRT true-score. equating. In A. A. von Davier & M. Liu (Eds.), *Population invariance of test equating and linking: Theory extension and applications across exams.* (ETS Research Report RR-06-31). Princeton, NJ: Educational Testing Service.

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.

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

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

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.

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.

Kim, S., Walker, M. E., & McHale, F. (2008). *Comparisons among designs for equating constructed response tests* (ETS Research Report RR-08-53). Princeton, NJ: Educational Testing Service.

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

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.

Liu, C., & Kolen, M. J. (2010). *A comparison among IRT equating methods and traditional equating methods for mixed-format tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, Colorado.

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

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1979). *A test of the adequacy of curvilinear score equating models*. Paper presented at the Computerized Adaptive Testing Conference, Minneapolis, MN.

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

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

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.

Sinharay, S., & Holland, P. W. (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.

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.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG (Version 7.03) [Computer software]. Lincolnwood, IL: Scientific Software International.

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.

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.

Table 1

Unweighted MC Moments in Original, 80% Response, and Imputed Data

Data	N	M	SD	Skew	Kurt
Original	301,095	29.37	12.05	-.25	-.58
80% Response	247,197	31.77	11.36	-.45	-.21
Imputed	247,197	38.12	9.48	-.74	.12

Table 2

Distribution of Parental Education in Groups of Varying ESs

Group	Target ES	Obs ES	% Cat1	% Cat2	% Cat3	% Cat4
New	0	.03	.12	.20	.36	.32
Old			.12	.20	.36	.32
New	.25	.24	.16	.27	.48	.09
Old			.04	.04	.49	.43
New	.50	.44	.26	.44	.29	.01
Old			.01	.01	.29	.69
New	.75	.72	.85	.09	.03	.03
Old			.03	.03	.09	.85

Table 3

Composite Score Moments for Parental Education Subgroups

Parent ED Level	Moments	UEq	SEq	IRT True	IRT Obs
Total	Mean	125.810	125.806	125.832	125.808
	SD	29.472	29.484	29.630	29.599
1	Mean	104.629	104.629	104.800	104.695
	SD	31.379	31.368	31.408	31.478
2	Mean	114.992	114.990	115.138	115.058
	SD	28.757	28.752	28.726	28.754
3	Mean	127.356	127.364	127.399	127.374
	SD	28.062	28.034	28.141	28.119
4	Mean	135.507	135.518	135.487	135.485
	SD	28.286	28.244	28.455	28.397

Note. UEq = Unsmoothed Equipercentile, SEq = Smoothed Equipercentile.

Table 4

REMSD Values for Parental Education Subgroups

Parent ED Level	UEq	SEq	IRT True	IRT Obs	SDTM
1	.047	.039	.019	.019	.017
2	.034	.029	.017	.014	
3	.034	.030	.010	.009	
4	.027	.022	.011	.010	

Table 5

Old Form Cut Scores and Cumulative Proportions

	Operational		Form B	
AP Grade	Cut Score	% Below	Cut Score	% Below
2	46.5	11.28	86.5	11.20
3	73.5	41.74	121.5	41.27
4	93.5	73.10	144.5	72.60
5	109.5	91.31	163.5	90.93

Table 6

New Form Cut Scores

		Parental Education Subgroup				
Method	Grade	Total	1	2	3	4
UEq	2	88.5	90.5	89.5	87.5	89.5
	3	123.5	123.5	123.5	122.5	123.5
	4	145.5	145.5	145.5	145.5	144.5
	5	163.5	163.5	163.5	163.5	164.5
IRT True	2	89.5	89.5	89.5	89.5	89.5
	3	123.5	123.5	123.5	123.5	123.5
	4	145.5	145.5	144.5	145.5	145.5
	5	163.5	163.5	163.5	163.5	163.5
IRT Observed	2	89.5	89.5	89.5	89.5	89.5
	3	123.5	123.5	123.5	123.5	123.5
	4	145.5	145.5	145.5	145.5	145.5
	5	163.5	163.5	163.5	163.5	163.5

Table 7

Classification Consistency for Parental Education Subgroups

Parent ED Level	UEq	IRT True	IRT Obs
1	99.01	100	100
2	99.51	98.68	100
3	98.20	100	100
4	97.51	100	100

Table 8

Old Form Equivalent Composite Score Moments

ES	Moments	SFE	SCE	IRT True	IRT Obs
0	Mean	124.146	124.024	124.159	124.171
	SD	30.794	30.790	31.549	31.445
.25	Mean	124.224	122.468	122.698	122.663
	SD	29.213	29.622	29.014	29.180
.50	Mean	123.597	119.809	120.839	120.763
	SD	30.440	32.181	30.573	30.682
.75	Mean	116.284	110.655	110.836	110.664
	SD	32.091	33.000	30.220	30.646

Note. SFE=Smoothed Frequency Estimation, SCE=Smoothed Chained Equipercentile.

Table 9

REMSD Values across ES Levels

ES	SFE	SCE	IRT True	IRT Obs	SDTM
.25	.124	.086	.095	.087	
.50	.230	.132	.152	.145	.017
.75	.333	.163	.221	.206	

Table 10

Classification Consistency with ES=0 as the Criterion

ES	SFE	SCE	IRT True	IRT Obs
.25	89.48	91.29	95.44	94.12
.50	75.18	84.37	86.68	85.36
.75	70.11	86.97	89.66	91.59

Table 11

Classification Consistency with IRT True Score Equating as the Criterion

ES	SFE	SCE	IRT Obs
0	98.68	97.34	98.68
.25	95.36	94.99	100
.50	89.82	97.00	100
.75	75.39	94.95	99.38

Table 12

Evaluation of FE and CE Equating Assumptions

ES	FE	CE
0	16.138	.040
.25	26.702	.084
.50	42.059	.146
.75	66.389	.158

Table 13

Form A Eigenvalues and Percentage of Variance Explained for the First Three Principal Components

		ES			
PC		0	0.25	.50	.75
1	Eigen	11.943	11.099	12.003	12.363
	%Var	27.770	25.810	27.910	28.750
2	Eigen	1.667	1.747	1.676	1.766
	%Var	3.880	4.060	3.900	4.110
3	Eigen	1.396	1.335	1.388	1.410
	%Var	3.250	3.110	3.230	3.280

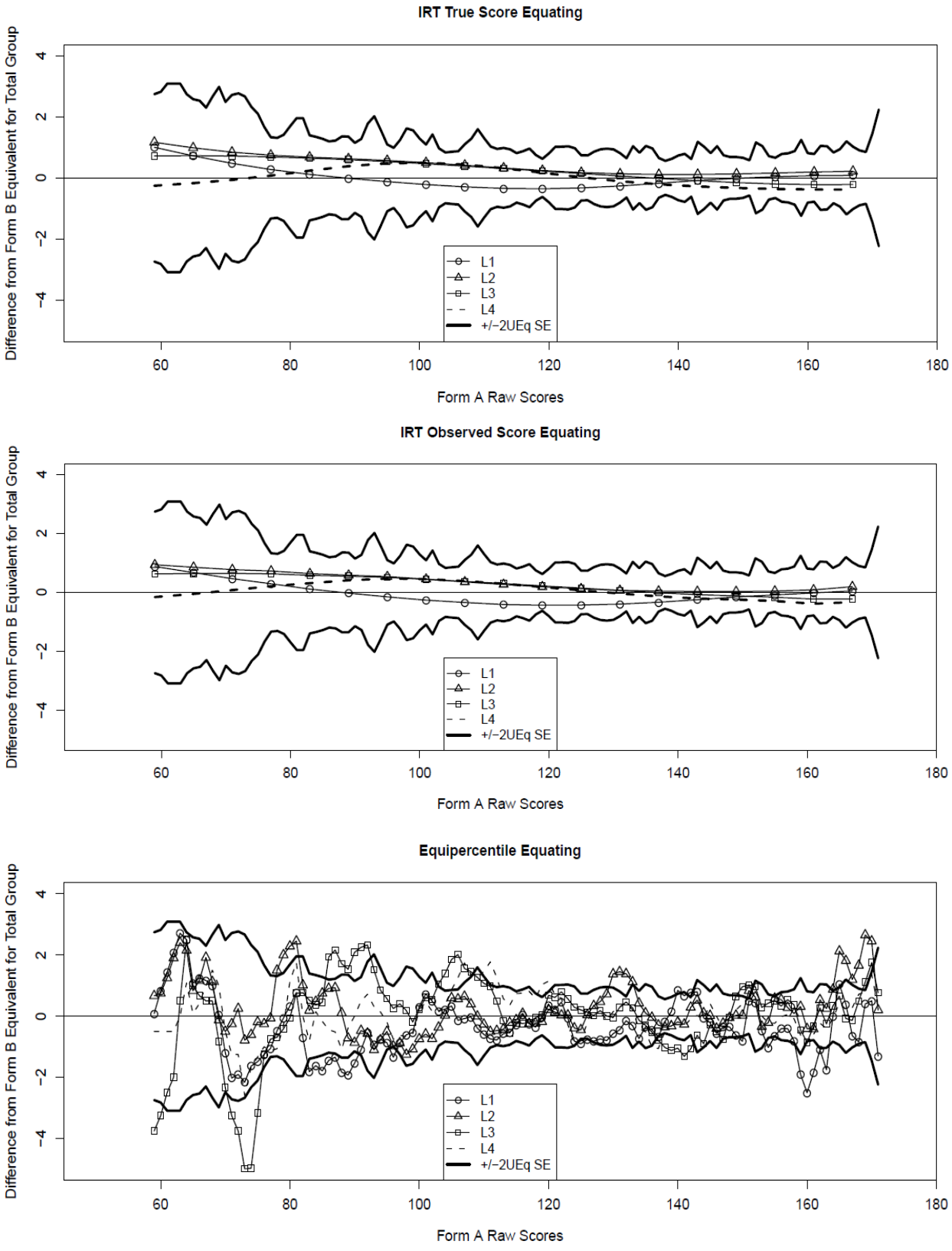

Figure 1. Comparison of SG equating relationships for parental education subgroups.

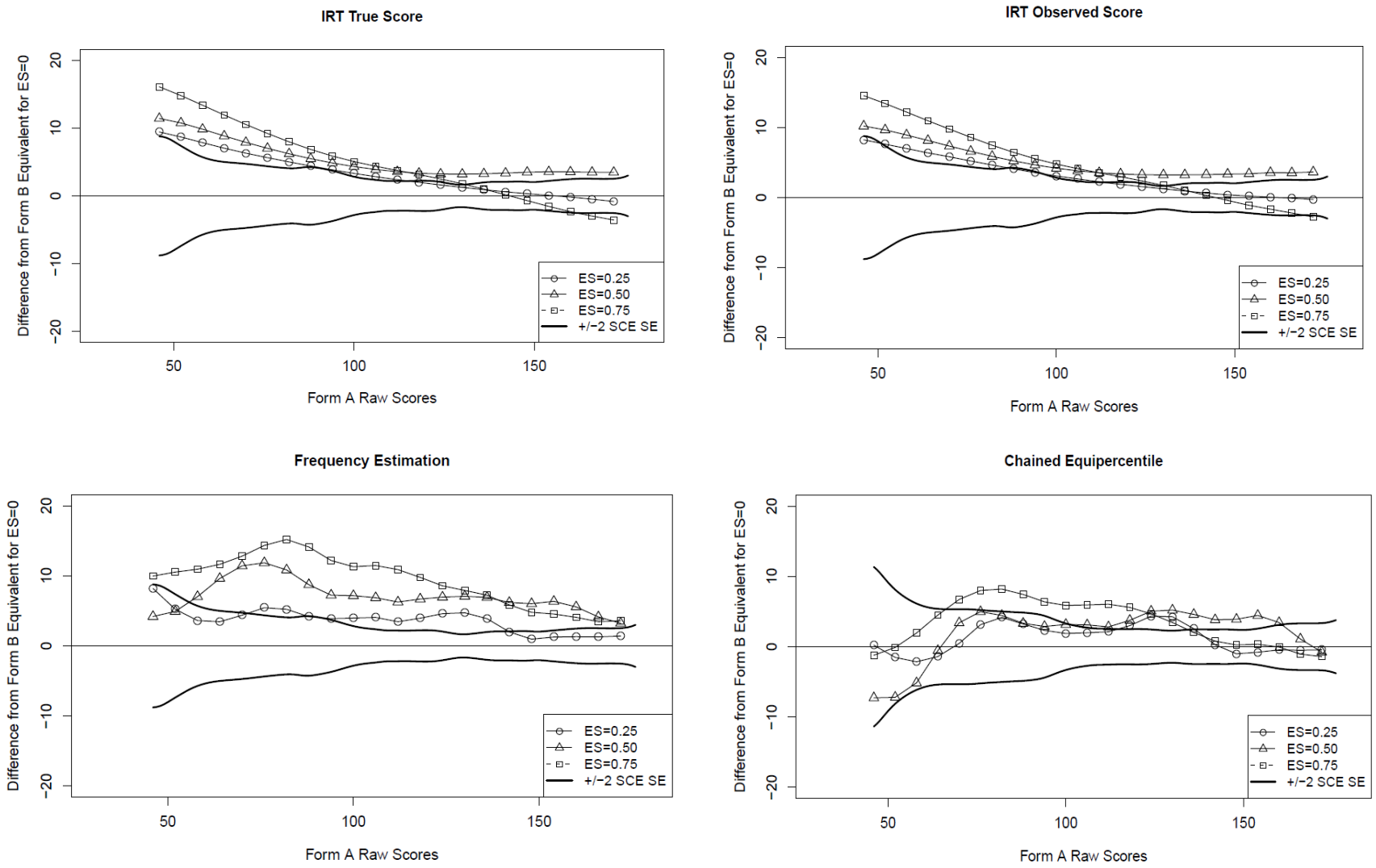

Figure 2. Comparison of IRT and traditional equating relationships for four ES levels.

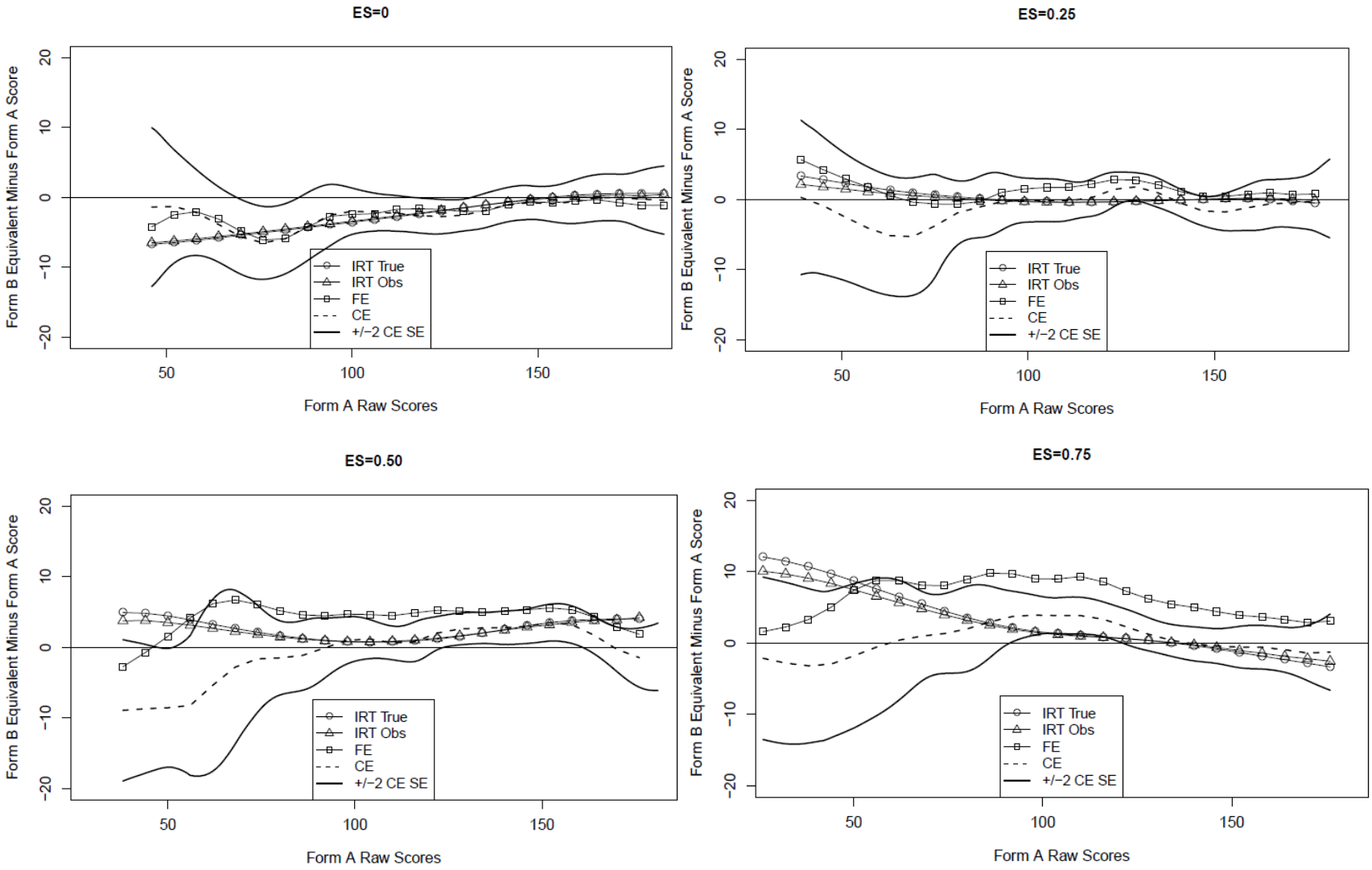

Figure 3. Comparison of equating relationships for four equating methods.

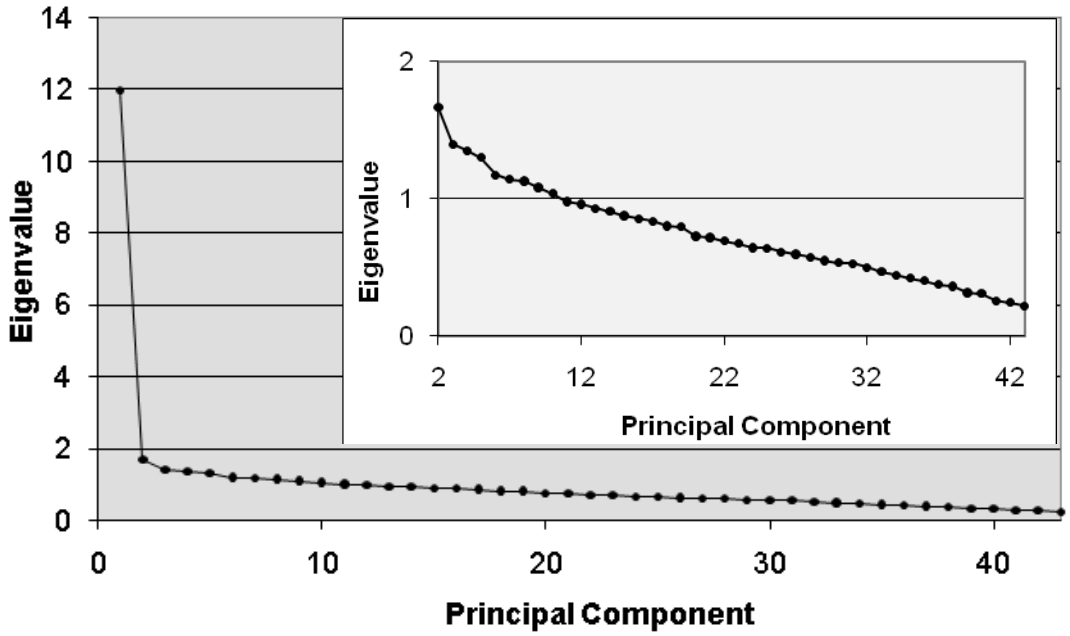

Figure 4. Scree Plots for Form A where ES=0.

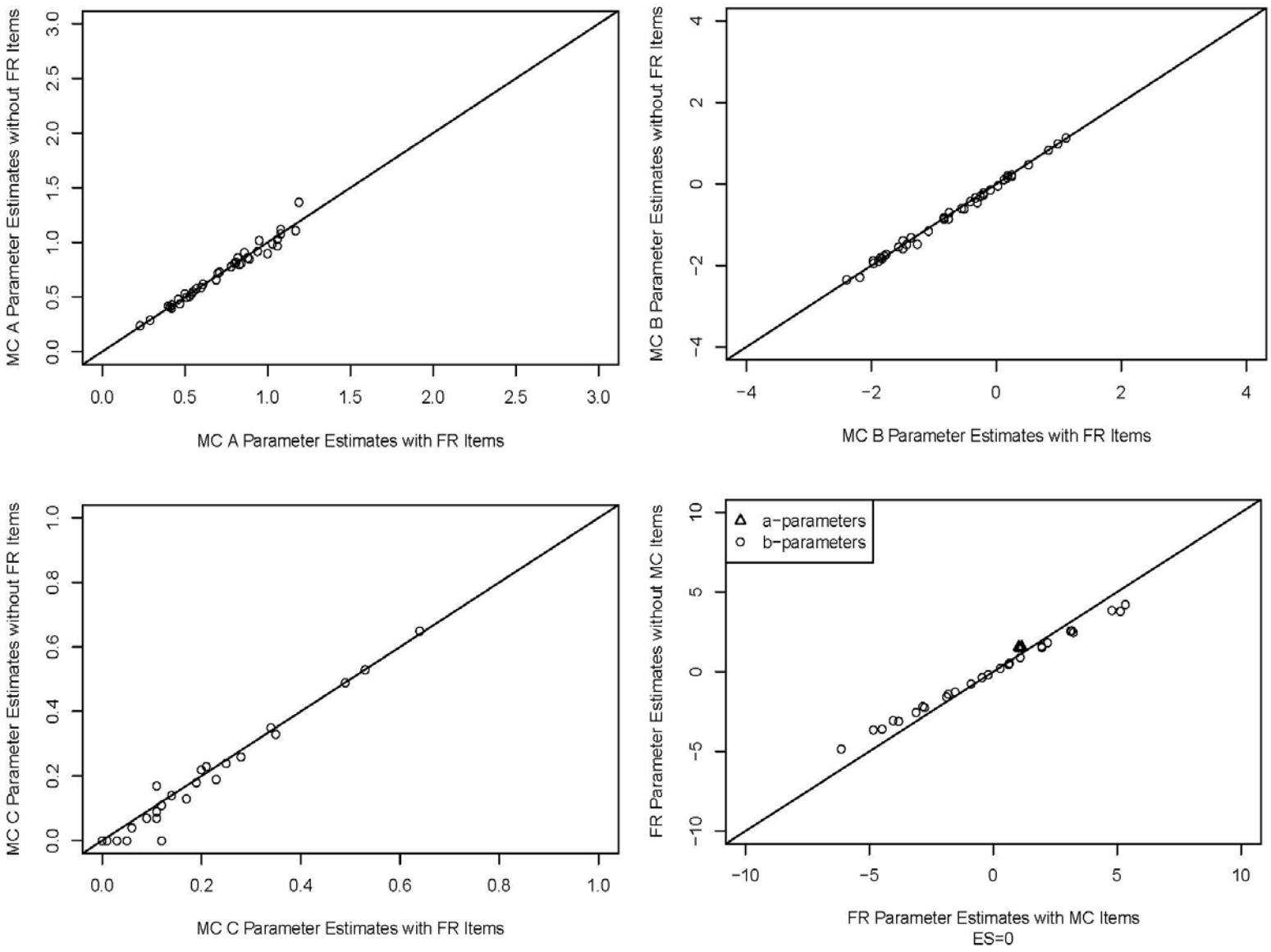

Figure 5. Comparison of Form A MC item parameter estimates for ES=0.

Chapter 7: Equity and Same Distributions Properties for Test Equating

Yi He and Michael J. Kolen

The University of Iowa, Iowa City, IA

Abstract

The purpose of this study was to evaluate how equating properties were preserved for mixed-format tests under the common item nonequivalent groups (CINEG) design. Real data analyses were conducted on 22 equating linkages of 39 mixed-format tests from the Advanced Placement (AP) Examination program. Four equating methods were used: the frequency estimation (FE) method, the chained equipercentile (CE) method, IRT true score equating, and IRT observed score equating. In addition, cubic spline postsmoothing was used with the FE and CE methods. The factors of investigation were the correlation between multiple-choice (MC) and free response (FR) scores, the proportion of common items, the proportion of MC-item score points, and the similarity between alternate forms. Results were evaluated using three equating properties: first-order equity, second-order equity, and the same distributions property. The main findings from this study were: (1) Between the two IRT equating methods, true score equating better preserved first-order equity than observed score equating, and observed score equating better preserved second-order equity and the same distributions property than true score equating; (2) between the two traditional methods, CE better preserved first-order equity than FE, but in terms of preserving second-order equity and the same distributions property, CE and FE produced similar results; (3) smoothing helped to improve the preservation of second-order equity and the same distributions property; (4) a higher MC-FR correlation was associated with better preservation of the equating properties for IRT methods; and (5) the proportion of common items, the proportion of MC score points, and the similarity between forms were not found to be associated with the preservation of the equating properties. These results are interpreted in the context of research literature in this area and suggestions for future research are provided.

Equity and Same Distributions Properties for Test Equating

An increasing number of testing programs have been using mixed-format tests, which typically contain dichotomously-scored multiple-choice (MC) items and dichotomously- or polytomously-scored free response (FR) items. The use of multiple formats presents a number of measurement challenges, one of which is how to accurately equate mixed-format tests under the common-item nonequivalent groups (CINEG) design. The CINEG design uses a common-item set to disentangle group differences from form differences and thus requires this common-item set to be representative of the entire test in terms of content and statistical properties (Kolen & Brennan, 2004). For mixed-format test equating under the CINEG design, ideally, the common items should contain both MC and FR items. However, in practice, for security and practical reasons, the common-item set often contains MC items only. Under these circumstances, how adequate is equating?

A number of studies have been conducted on mixed-format test linking and equating, focusing on the following aspects: (a) the impact of the characteristics of the common-item set (such as the common-item length, the representativeness) and the format effect (such as the correlation between MC and FR scores, and the relative proportion of MC and FR items/score points) on linking or equating accuracy (Bastari, 2000; Cao, 2008; Kim & Kolen, 2006; Kim & Lee, 2006; Lee, Hagge, He, Kolen, & Wang, 2010; Li, Lissitz, & Yang, 1999; Paek & Kim, 2007; Tan, Kim, Paek, & Xiang, 2009; Tate, 2000; Wu, Huang, Huh, & Harris, 2009); (b) the effectiveness of using MC items only as the common-item set (Kim & Lee, 2006; Kirkpatrick, 2005; Walker & Kim, 2009; Wu et al., 2009); (c) the impact of group differences or form differences on linking or equating accuracy (Bastari, 2000; Kim & Kolen, 2006; Kim & Lee, 2006; Lee et al., 2010; Li et al., 1999; Wu et al., 2009); and (d) the effect of multidimensionality (Cao, 2008; Kim & Kolen, 2006). Generally, these studies revealed that a stronger correlation between MC and FR scores, a higher proportion of MC score points, a longer common-item set, and more similar groups, led to more effective linking or equating.

The purpose of this study was to evaluate how equating properties are preserved when the

data were collected under the CINEG design and the common-item set contains MC items only. Specifically, this study investigated whether and how the following factors affect the preservation of equating properties: (a) the correlation between MC and FR scores; (b) the proportion of common items; (c) the proportion of MC-item score points; (d) the similarity between forms; and (e) equating methods. The equating methods used in this study include the frequency estimation (FE) method, the chained equipercentile (CE) method, IRT true score equating, and IRT observed score equating. Three equating properties were used as evaluation criteria: first-order equity, second-order equity, and the same distributions property.

Equating Properties

The concept of the *equity property* of equating was first proposed by Lord (1980, p. 195). Lord's equity property holds only if it is a matter of indifference for an examinee whether he or she is administered the old form or the new form. In other words, if for examinees with a given true score, the distribution of the equated scores on the new form is identical to the distribution of the scores on the old form, then Lord's equity property holds.

Lord's equity property will not hold unless the two forms are identical, in which case equating is unnecessary (Kolen & Brennan, 2004). For this reason, Divgi (1981) and Morris (1982) proposed the concept of the *weak equity property* or *first-order equity property* (also referred to as *tau equivalence* by Yen, 1983), which is less restrictive than Lord's equity property. First-order equity requires that the expected conditional means are equal for the alternate forms after equating. Kolen and Brennan (2004, p. 301) also described the *second-order equity property*, which holds if the conditional standard errors of measurement (SEM) are equal for the alternate forms after equating.

The same distributions property (Tong & Kolen, 2005; also called *observed score equating property* in Kolen & Brennan, 2004, or *equipercentile equating property* in Kim, Brennan, & Kolen, 2005) is considered to hold if scores on the equated new form have the same distribution as scores on the old form.

Procedures have been developed to assess these properties. Kolen, Hanson, and Brennan

(1992) described procedures to calculate the conditional means and SEMs under strong true score models. Kolen, Zeng, and Hanson (1996) also presented procedures to find the conditional means and SEMs under dichotomous IRT models; Wang, Kolen, and Harris (2000) described procedures that can be implemented with polytomous IRT models. All three procedures can be used to assess the first- and second-order equity properties. Tong and Kolen (2005) and Kim et al. (2005) also created specific indices to assess the preservation of the equity properties.

To evaluate the preservation of the same distributions property, Tong and Kolen (2005) used the Kolmogorov-Smirnov *T* (Conover, 1999), which is the largest absolute difference between two distribution functions, and Kim et al. (2005) used the sum of absolute differences (EQAD), which is the absolute differences between two distributions summed over all score points.

The equity properties and the same distributions property were used in the literature to evaluate the effectiveness of linking and equating. Harris (1991) used the first-order equity criterion to compare Angoff's two data collection designs using equipercentile and IRT equating methods in a vertical scaling context. Bolt (1999) used the first-order and second-order equity criteria to evaluate the effect of multidimensionality on IRT true score equating results. Zhang (2007) used the criteria to compare IRT equating results based on the 3PL model versus based on the Generalized Partial Credit model in equating testlet-based tests.

Two studies (Kim et al., 2005; Tong & Kolen, 2005) are especially relevant to the current study, because they made use of the same distributions property and equity properties as criteria to compare equating methods. Tong and Kolen (2005) conducted a real data analysis as well as a simulation study to compare the equipercentile method, IRT true score equating, and IRT observed score under the random groups design. The factor of investigation was the similarity between alternate forms. They found that when alternate forms were similar, all three methods preserved the three equating properties well; when the alternate forms were different, the IRT true score method outperformed the other two methods in terms of preserving the first-order equity property; and both the IRT observed score equating method and the equipercentile method

outperformed the IRT true score method in preserving the same distributions property and second-order equity. In addition, both the real data analysis and the simulation study showed that greater similarity between alternate forms was associated with better preservation of second-order equity. Further, the simulation study showed that the more similar the alternate forms were, the better first-order equity was achieved.

Kim et al. (2005) used the three equating properties to compare two true score equating methods (beta 4 true score equating and IRT true score equating) and two observed score equating methods (beta 4 observed score equating and IRT observed score equating), also under the random groups design. They found that whether the beta 4 model or the IRT model was assumed, true score equating better preserved first-order equity than observed score equating, and observed equating outperformed true score equating in achieving second-order equity and the same distributions property.

These equating properties have not been used to compare equating methods under the CINEG design, nor to assess equating results for mixed-format tests. The current study was intended to further the understanding about how these properties are preserved in mixed-format test equating under the CINEG design.

Method

Data

The data used in the study were from the College Board's Advanced Placement (AP) exams. Each exam has two or more forms, administered in different years. Forms across years of the same test are linked by a set of common MC items. Hence the equating design used in the study was the CINEG design (Kolen & Brennan, 2004). Altogether, there are 22 pairs of alternate forms to be equated, i.e., 22 equating linkages. Table 1 provides the AP tests and forms used in the current study. The solid horizontal lines between tests/forms show equating linkages. Operationally, the MC items in the AP exams were formula scored, and each MC item and FR item received a non-integer weight. Formula scoring and non-integer weights resulted in non-integer scores, which could not be easily handled by currently available psychometric

programs. Thus, in the current study, responses were rescored using number-correct scoring and integer weights were assigned. The integer weights were determined to mirror the non-integer weights used in the operational situation as much as possible. That is to say, the proportions of points that each MC item and FR item contributed to the total score points in the current study were similar to the proportions that they contributed to the total score points operationally. The number of MC score points, number of FR score points, and the integer weights for each item are also provided in Table 1.

Another consequence of formula scoring is that a large number of MC items were not answered because examinees were instructed that if an MC item was incorrectly answered, some points would be deducted. In the current study, only examinees answering all the MC items were included in the analyses. The initial sample sizes (before eliminating examinees with missing responses) in the original data and the actual sample sizes used in the current study are provided in the third and fourth columns of Table 2. This table also presents the means and standard deviations for the composite scores and for the common-item scores, as well as effect sizes (effect size $= \frac{|\overline{X}_{old} - \overline{X}_{new}|}{\sqrt{(SD_{old}^2 + SD_{new}^2)/2}}$, where $\overline{X}_{old}$ and $\overline{X}_{new}$ are the means of common-item scores for the old form and the new form respectively, and SD_{old} and SD_{new} are the standard deviations of common-item scores for the old form and the new form respectively) (Yen, 1986) between the forms for each linkage. Generally, the group differences were small: 16 out of the 22 equating linkages had effect sizes below .1, and the other six had effect sizes ranging from .1075 to .3037.

Equating

Equating was conducted using two traditional equipercentile methods (FE and CE) and two IRT methods (IRT true score equating and IRT observed score equating). For the traditional methods, cubic spline postsmoothing procedures (Kolen & Brennan, 2004) were used. Smoothing parameters of .1 and .3 were selected following a procedure described in Kolen and Brennan (2004): Within the score range with percentile ranks of .5 and 99.5, the highest *S* value

that resulted in a smoothed equating relationship within error bands of plus and minus one raw score standard error was chosen. *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009) was used to conduct traditional equating.

IRT equating involves multiple steps. The first step is item calibration. The three-parameter logistic (3PL) model (Birnbaum, 1968) was assumed for the MC items and the graded response (GR) model (Samejima, 1969) was assumed for the FR items. Separate calibration was conducted for each form. PARSCALE (Muraki & Bock, 2003) was used for IRT calibration. Next, the Stocking-Lord method (Stocking & Lord, 1983) was used to transform the estimated item parameters and the quadrature points of the posterior distribution on the new form scale to the old form scale. The computer program STUIRT (Kim & Kolen, 2004) was used for scale transformation. Finally, with the estimated item parameters and the proficiency distributions on the same scale across forms, IRT equating was conducted using POLYEQUATE (Kolen, 2004b).

Note that the FE method and IRT observed score equating require the definition of a synthetic group. In such cases, the synthetic group was defined as the group of examinees taking the new form (Group 1).

Results Evaluation

To evaluate the similarity between forms and the same distributions property, the Kolmogorov-Smirnov *T* (Conover, 1999), which is a nonparametric statistic to compare two distributions, was used. Since the groups were non-equivalent, the comparison of the distributions was based on the synthetic group which was defined as the new form group (Group 1) in this study. The distributions of the old and new form scores for the synthetic group were obtained assuming the IRT models (3PL for the MC items and GR for the FR items). Let *X* denote the random variable score on the new form (Form X) and *Y* denote random variable score on the old form (Form Y). Denote the empirical distribution function (*edf*) for scores on Form X for Group 1 as $F_1(x)$, and the *edf* for scores on Form Y for Group 1 as $G_1(y)$. The Kolmogorov-Smirnov *T* (referred to as T_1 here) is defined as the largest vertical distance between

the two *edf*s graphically. Mathematically, T_1 is calculated as the maximum of the absolute differences between the two distributions for all $x = y$, i.e.,

$$T_1 = \sup_{x=y} | F_1(x) - G_1(y) | . \quad (1)$$

When the distributions of the two forms are identical, T_1 would be 0. The more different the two distributions are, the larger is the value of T_1.

To assess the preservation of the same distributions property, the Kolmogorov-Smirnov T (referred to as T_2 here) was used again. Define $eq_Y(x)$ as the equating function to convert scores on Form X to the scale of Form Y, and $e\hat{q}_Y(x)$ as the sample estimate of a score on Form X equated to Form Y. The index T_2 refers to the largest difference between the *edf* of the equated scores on Form X and the *edf* of the scores on Form Y for the synthetic group (Group 1) for all $e\hat{q}_Y(x) = y$, i.e.,

$$T_2 = \sup_{e\hat{q}_Y(x)=y} | F_1[e\hat{q}_Y(x)] - G_1[y] | . \quad (2)$$

The difference between T_1 and T_2 is that T_1 compares the distributions of scores on the old form and on the new form before equating, whereas T_2 compares the distributions after equating. Like T_1, the larger the T_2 value, the more disparate are the distributions of the old form and the equated new form.

The First- and second-order equity properties were assessed using the IRT framework. First-order equity was evaluated using the index D_1 (Tong & Kolen, 2005), which is calculated by

$$D_1 = \frac{\sqrt{\sum_i q_i \{E[Y | \theta_i] - E[e\hat{q}_Y(x) | \theta_i]\}^2}}{SD_Y}, \quad (3)$$

where $E[Y | \theta_i]$ is the old form conditional mean for a given proficiency θ_i, $E[e\hat{q}_Y(x) | \theta_i]$ is the conditional mean of equated score for a given proficiency θ_i, q_i is the quadrature weight at θ_i, and SD_Y is the standard deviation of Form Y. The quadrature points θ_i and the

corresponding quadrature weight q_i can either come from the examinees taking the old form or those taking the new form, depending on the definition of the synthetic population. In this study, because the synthetic group was conceptualized as the examinees taking the new form, the quadrature points which had been transformed in the scale transformation step and the corresponding quadrature weights for the examinees taking the new form were used. The smaller the D_1 value, the better first-order equity is preserved. The denominator SD_Y in the equation is used to standardize the index so that D_1 indices from different tests can be compared.

Second-order equity was evaluated using the index D_2 (Tong & Kolen, 2005), which is calculated by:

$$D_2 = \frac{\sqrt{\sum_i q_i (SEM_Y \mid \theta_i - SEM_{\hat{e}q_Y(x)} \mid \theta_i)^2}}{SD_Y}, \qquad (4)$$

where $SEM_Y \mid \theta_i$ is the conditional SEM for the old form for examinees with proficiency θ_i, and $SEM_{\hat{e}q_Y(x)} \mid \theta_i$ is the conditional SEM for the equated new form for examinees with proficiency θ_i. Similar to the D_1 index, the quadrature points and weights from the examinees taking the new form were used in this study. A large D_2 value suggests that the second-order equity property is not sufficiently preserved. The denominator SD_Y in the equation is used to standardize the index so that D_2 indices from different tests can be compared.

The conditional means and conditional SEMs, which are required for the calculation in Equations 3 and 4, were obtained following the procedure described in Wang et al. (2000). The computer program POLYCSEM (Kolen, 2004a) was used for the implementation of the procedure. D_1 and D_2 indices were then calculated for each equating method.

Spearman correlation coefficients were calculated to address the questions whether there are relationships between the preservation of the equating properties and factors such as the MC-FR correlation, the proportion of common items relative to the total number of MC items, the proportion of MC score points relative to composite score points, and the similarity between

forms. Because for each equating linkage there were two MC-FR correlations, two proportions of common items, and two proportions of MC-item score points, one from the old form and the other from the new form, to calculate the Spearman correlation coefficients, the average of the old form value and the new form value was used. Table 3 presents the values for these factors. Also, because three exams have double linkages, i.e., Environmental Science 2004 was linked to 2006 and 2007, Spanish Literature 2004 was linked to 2005 and 2006, and World History 2004 was linked to 2005 and 2006, to avoid these dependence issues, three equating linkages (Environmental Science 2004-2007, Spanish Literature 2004-2005, and World History 2004-2005) were excluded from the calculation of the Spearman correlations.

Results and Discussions

Evaluation of Dimensionality

Throughout the study, unidimensional IRT models were heavily used, so it is important to assess if the unidimensionality assumption held. Two methods were used to evaluate dimensionality. The first was to check the disattenuated correlation between MC and FR. The disattenuated correlation represents the correlation between the true scores of two variables. If the disattenuated correlation is close to 1, the MC items and the FR items are essentially measuring the same construct, and multidimensionality due to item formats is less likely. The second method was to conduct a principal components analysis on the matrix of inter-item tetrachoric (for MC items) or polychoric (for FR items) correlations. The reason that tetrachoric and polychoric correlations were used instead of conventional Pearson correlations is that the use of Pearson correlation matrices might overestimate the number of true underlying dimensions by including the difficulty factor as additional dimensions (Hambleton & Rovinelli, 1986; McDonald & Ahlawat, 1974; Wherry & Gaylord, 1944), whereas with tetrachoric and polychoric correlations, the issue with the confounding difficulty factors is much less severe. SAS software, Version 9.2 of the SAS System for Windows, was used to obtain tetrachoric and polychoric correlations and to perform the principal components analysis. The criteria for unidimensionality include: (a) An “elbow” occurs at the second eigenvalue in the scree plot; (b) the ratio of the first

to second eigenvalues is large relative to the ratios of the second to third and the third to fourth eigenvalues (Lord, 1980); and (c) the first eigenvalue accounts for at least 20% of the total variance (Reckase, 1979).

For most of the exams, there was no strong evidence against unidimensionality, regardless of the different methods and different criteria used to assess dimensionality. For some exams, there was some evidence suggesting multidimensionality. The determination of multidimensionality varies, depending on the evaluation method/criteria. All these methods agreed that Spanish Literature was subject to multidimensionality. In addition, the disattenuated correlations suggested that the English Language and Composition forms were multidimensional. The principal components analyses suggested that the Spanish Language and Environmental Science forms were multidimensional. See He (2011) for a complete set of tables and figures regarding dimensionality. Although multidimensionality was detected for some exams, it still was considered worthwhile to observe what happens when the methodology is applied to tests that might not be fit well with a unidimensional IRT model. However, it should be noted that the conclusions for these exams are limited.

Similarity between Alternate Forms and the Same Distributions Property

Table 4 lists the Kolmogorov-Smirnov T_1 and T_2 values for the 22 equating linkages. The T_1 values ranged from .014 to .198. Five equating linkages had Kolmogorov-Smirnov T_1 values over .1, ten between .05 and .1, and seven lower than .05. Among the 22 equating linkages, the Spanish Language and Spanish Literature exams tended to show more form differences than the other exams.

Columns 3 to 10 in Table 4 provide the Kolmogorov-Smirnov T_2 values for the 22 equating linkages and for the various equating methods used in the study. For both the IRT equating methods, the T_2 values were noticeably lower than T_1, especially for those exams which had large T_1, indicating that IRT equating methods effectively reduced form differences, and thus preserved the same distributions property well. In addition, it appears that the T_2 values with IRT observed score equating were lower than those with IRT true score equating. This is consistent

with the results in Kim et al. (2005) and Tong and Kolen (2005).

For the traditional equating methods, the T_2 values were lower than T_1 for most of the exams. However, for Biology 2004-2006, Chemistry 2004-2006, and German Language 2005-2007, the T_2 values were even larger than the T_1 values, for both FE and CE, whether smoothed or not. For European History 2005-2007, the T_2 value with the unsmoothed FE method was larger than the T_1 value. This means that after equating, the score distributions became more discrepant.

There are two possible explanations for this result. First, the Kolmogorov-Smirnov T statistics were calculated based on the distributions of the synthetic group, which were obtained using IRT which might have biased the results for the traditional methods. Second, when the distributions of the scores on the old and new forms are already very similar before equating (here for Biology 2004-2006, Chemistry 2004-2006, and German Language 2005-2007, the T_1 values were around .02 or lower), equating might make the distributions more discrepant due to equating error.

Between the FE and CE methods, one method did not appear to outperform the other in terms of preserving the same distributions property: Half of the equating linkages had lower T_2 values with the unsmoothed FE method, and half had opposite results. The T_2 values with smoothing tended to be lower than those without smoothing, for both the FE and CE methods. This finding suggests that smoothing makes the equated new form distribution more similar to the old form distribution, thus helping to preserve the same distributions property.

Figure 1 shows the score distributions of the old and new forms before and after equating for Spanish Literature 2004-2006. It contains nine plots: The upper left plot illustrates the score distributions of the two forms before equating, and the other eight compare the score distributions of the old form and the equated new form for each of the eight equating methods. For all nine plots, the score distributions of the two forms were based on the synthetic group. In each plot, the curve represented by the dashed line represents the cumulative distribution of the scores on the old form, and the curve represented by the solid line represents the cumulative

distribution of the scores on the new form. Thus, the dashed lines are the same across all nine plots, and the solid lines are different before equating and after equating, and are different for different equating methods. Consistent with the findings from the Kolmogorov-Smirnov *T* statistics, the general tendency was that the score distributions of the old and new forms after equating were closer to each other than those before equating. For IRT equatings, the two curves representing the cumulative frequencies for the old and new forms were almost on top of each other; for traditional equatings, although the score distributions of the old and new forms after equating were not as close as for IRT equatings, they were still closer than before equating. Similar patterns were found for other equating linkages, except for Biology 2004-2006, Chemistry 2004-2006, and German Language 2005-2007, where the distributions were less close after traditional equatings than before equating. See He (2011) for a complete set of figures.

First-Order Equity

Table 5 provides the D_1 values for the various methods and equating linkages used in this study. Median values for each column are also presented. Between the two traditional methods, the CE method appears to better preserve the first-order equity property than the FE method: For 19 out of the 22 equating linkages, the D_1 values with CE were lower than those with FE. The three exceptions were Chemistry 2005-2007, German Language 2004-2006, and Spanish Literature 2004-2005. In addition, smoothing does not seem to lower the D_1 value, suggesting that smoothing did not help improve the preservation of first-order equity.

Between the two IRT methods, IRT true score equating appears to produce lower D_1 values than the observed score method, except for three equating linkages (European History 2005-2007, German Language 2004-2006, and World History 2004-2005). This is consistent with the previous research findings (Kim et al., 2005; Tong & Kolen, 2005) that the IRT true score method outperformed the IRT observed score method in terms of preserving first-order equity under the random groups design. Comparing the traditional with the IRT methods, it seems that the IRT methods led to much lower D_1 values than the traditional methods. This might be due to the criterion being based on the IRT model, and thus the criterion might favor the IRT

equating methods.

The D_1 index empirically quantifies the overall difference in expected scores between the old and the new forms across all the proficiency levels. To further compare how well first-order equity holds conditionally, first-order deviation was assessed graphically. Figure 2 provides an example for French Language 2004-2006. In Figure 2 there are four plots comparing: (a) the FE method versus the CE method; (b) IRT true score equating versus IRT observed score equating; (c) smoothing versus non-smoothing for the FE method; and (d) smoothing versus non-smoothing for the CE method. In each plot, the horizontal axis represents scores on the old form, and the vertical axis represents standardized difference, which is the difference between in the expected scores on the two forms divided by the standard deviation of the old form. Thus each unit on the vertical axis is one SD of difference. If first-order equity had been preserved perfectly, there would have been no difference between the expected scores on the two forms at each proficiency level, and the curves in each plot would have been a straight horizontal line with a value 0 at the vertical axis (the dot dash line in each plot). The closer the curves are to the dot dash line, the better first-order equity holds.

In Figure 2, plot (a) does not seem to show that the curve for one method was consistently closer to the straight horizontal line than the curve for the other method. In plot (b), it appears that the curve for IRT observed score equating deviated slightly more from the horizontal line than the curve for IRT true score equating, suggesting that first-order equity was violated more by the IRT observed score equating method, and the difference in the extent of violation was very small. In plots (c) and (d), for either the FE or the CE methods, generally, the curves with smoothing did not appear to be closer to the horizontal line than the curves without smoothing, and the curves with a higher degree of smoothing (S = .3) were not closer to the horizontal line than those with a lower degree of smoothing (S = .1). Plots for other equating linkages show similar patterns. See He (2011) for a complete set of figures.

Second-Order Equity

Table 6 provides the D_2 values for the various methods and equating linkages used in this

study. Median values for each column are also presented. Between the two traditional methods, the D_2 values were close. It does not appear that the D_2 values with one method were consistently lower than those with the other method. Therefore, it is difficult to conclude which method performs better in terms of preserving second-order equity. In addition, smoothing tended to lower the D_2 values, and a higher smoothing degree (S = .3) tended to result in lower D_2 values than a lower smoothing degree (S = .1). Between the two IRT methods, IRT observed score equating tended to have lower D_2 values than the true score method, which is consistent with the previous research findings (Kim et al., 2005; Tong & Kolen, 2005) that the IRT observed score equating method outperforms the IRT true score equating method in preserving second-order equity.

The D_2 index empirically quantifies the overall difference between the old and the new forms in conditional standard errors of measurement (SEM) across all the proficiency levels. To further compare how well second-order equity holds conditionally, second-order deviation was assessed graphically. Figure 3 provides an example for European History 2005-2007. It contains four plots comparing: (a) the FE method versus the CE method; (b) IRT true score equating versus IRT observed score equating; (c) smoothing versus non-smoothing for the FE method; and (d) smoothing versus non-smoothing for the CE method. In each plot, the horizontal axis represents scores on the old form, and the vertical axis represents the difference in the conditional SEMs across the two forms. If the second-order equity property held perfectly, then there would have been no difference between the conditional SEMs on the two forms at each proficiency level, and the curves in each plot would have been a straight horizontal line with a value 0 at the vertical axis (the dot dash line in each plot). The closer the curves were to the dot dash line, the better the second-order equity property held.

In plot (a) of Figure 3, it does not seem that the curve for one traditional equating method was consistently closer to the horizontal line than the curve for the other method. In plot (b), the curves representing the two IRT methods were very close to each other. It appears that in the middle part of the score range, the curve for IRT true score equating deviated slightly more from

0 than the curve for IRT observed score equating, suggesting that the second-order equity property was more violated by the IRT true score method in this part of the score range. In plots (c) and (d), for the majority part of the score points, the three curves were quite close to one another. It appears that for both plots (c) and (d), the curves with $S = .3$ were slightly closer to 0 than the curves with unsmoothed methods or with $S = .1$. In addition, the curves with $S = .3$ tended to be the smoothest, and the curves with the unsmoothed methods tended to be the bumpiest. Plots for other exams show similar patterns. See He (2011) for a complete set of figures.

Table 7 provides the Spearman correlation coefficients between each of the indices measuring the preservation of the three equating properties and each of the various factors of investigation, along with the corresponding *p* values. The proportion of common items, the proportion of MC score points, and T_1 do not appear to be associated with the preservation of the three equating properties: The Spearman correlation coefficients were not significantly different from 0, with *p* values far above the .05 level, for both the traditional and IRT methods. This finding is not consistent with Tong and Kolen (2005) that more similar alternate forms led to better preservation of the equity properties. One possible explanation is that the tests used in the current study did not have as much form difference as the tests used in their study did. The magnitude of the correlation coefficients could decease when the range of the variables is smaller. Thus, it is possible that if the current study had tests with larger differences in alternate forms, relationships between the similarity of forms and the preservation of the equating properties might be found.

Some relationships with correlations that were significant from 0 at the .05 level were observed between the MC-FR correlation and the D_1, D_2, and T_2 values for the IRT methods: Between the MC-FR correlation and D_1, the coefficients were -.68 with *p*-values of .0014 for both IRT true score equating and IRT observed score equating; between the MC-FR correlation and D_2, the Spearman correlation was $-.49$ with a *p*-value of .0320 for IRT true score equating, and the Spearman correlation was $-.37$ with a *p*-value of .1226 for IRT observed score

equating; and between the MC-FR correlation and T_2, the Spearman correlation coefficient was .45 with a *p*-value of .05 for IRT observed score equating.

Overall, limited relationships were found between the factors (the MC-FR correlation, the proportion of common items, and the proportion of MC score points, and the similarity of alternate forms) and the preservation of the three equating properties. Given the low variability in these factors, it is possible that there might have been more significant relationships between any of these factors and the D_1, D_2, and T_2 values had the variability been larger.

Conclusions and Practical Implications

Overall, the results from this study provided evidence supporting that the equity properties and the same distributions property could be reasonably preserved when equating mixed-format tests under the CINEG design using only MC items as the anchor. These findings are likely due, in part, to the similarity of the test forms that were equated in this study as well as to the similarity of the examinee groups used in each equating that was conducted.

The same distributions property and second-order equity tended to be better satisfied with methods that are based on observed score distributions, such as the IRT observed score method (as opposed to true score equating), smoothing (as opposed to no smoothing), and a higher degree of smoothing (as opposed to a lower degree of smoothing), while first-order equity appeared to be better satisfied with methods that are based on true score distributions, such as IRT true score equating (as opposed to observed score equating). Between the FE and CE methods, because both methods are based on observed score distributions, it is difficult to conclude which method performs better relative to any of the three criteria from the current analysis. The finding that IRT true score equating outperformed IRT observed score equating in terms of preserving first-order equity, and IRT observed score equating was better than IRT true score equating in preserving the same distributions property and second-order equity is consistent with Kim et al. (2005) and Tong and Kolen (2005), although the latter two studies were under a different equating design (i.e., the random groups design).

A higher MC-FR correlation tended to be associated with better preservation of

first-order equity for both IRT methods, with better preservation of second-order equity for IRT true score equating, and with better preservation of the same distributions property for IRT observed score equating. The proportion of common items, the proportion of MC-item score points, and the similarity between forms do not appear to be related to the preservation of the equating properties.

Taking the findings from the current study and the previous studies together, some suggestions about the choice of equating methods could be made. Generally, first-order equity focuses on test score comparability at the level of the individual examinee, second-order equity investigates measurement precision, and the intent of the same distributions property is to maintain test score comparability with respect to the entire group of examinees. Operationally, preserving the three properties should be prioritized according to the specific focus of the testing program.

For a testing program that routinely employs IRT models to construct test forms and that uses IRT equating methods, if the purpose is to prevent individual examinees from being advantaged or disadvantaged by being administered one test form versus another, then satisfying first-order equity might be of primary concern. In this case, the IRT true score equating method might be recommended over the observed score equating method. If the priority is to achieve the same percentages of examinees across alternate forms at each cut score, then satisfying the same distributions property might be considered to be of primary importance, and IRT observed score equating might be preferred over true score equating. If the focus is on improving measurement precision and decreasing measurement errors, then preserving second-order equity is an important criterion, as well, and IRT observed score equating might be used instead of IRT true score equating.

The findings from this study suggest that when the test forms are mixed-format but the common-item set contains MC items only, increasing the correlation between MC scores and FR scores might help to better achieve the equating properties if IRT equating methods are used. In practice, a realistic way to increase the MC-FR correlation is to improve the reliability of scores

for each item format.

In a testing program where test forms are constructed based on classical test theory and traditional equating methods are used, CE might be preferred over FE if the primary interest is to prevent individual examinees from being advantaged or disadvantaged by being administered one test form versus another. In addition, for both the CE and FE methods, smoothing is recommended since it might enhance the preservation of second-order equity and the same distributions property. Although a higher degree of smoothing tends to help better satisfy second-order equity and the same distributions property, smoothing parameters should be carefully selected since a higher degree of smoothing also introduces more systematic error (Kolen & Brennan, 2004).

This study examined how first- and second-order equity held for equated raw scores. Practitioners in testing programs should also check the equity properties for scale scores. In particular, practitioners should carefully inspect how first-order equity is achieved at the score points where cut scores are located. The Difference That Matters (DTM; Dorans & Feigenbaum, 1994) could be used as a criterion to determine the acceptable range of the deviation from first-order equity (the difference between the expected means on the old form and the equated new form). The DTM is defined as half of the reporting score unit. The rationale of the DTM is that if two unrounded scale scores are within half a unit of each other, they may ultimately become the same reported score after rounding. Thus, a difference of less than one-half reported score unit may be considered acceptable. However, a difference of more than one-half score unit might lead to a change in the reported score. In the context of first-order equity, if the deviation from first-order equity is more than half a unit of the reported scale score (especially at around a cut score), it should probably be considered of concern, because this amount of deviation may lead to falsely passing examinees who should have failed or falsely failing examinees who should have passed. Of course, as Kolen and Brennan (2004) pointed out, the DTM criterion is a guideline rather than a rigid rule. Also, at the score range where there are very few examinees, first-order deviations larger than a DTM might not be a serious problem.

Limitations and Future Research

This study used empirical data. One advantage of using empirical data is that they reflect how real data behave in practice. However, with empirical data, because many factors are interacting with each other, conclusions might only be made within the scope of the particular data set used for the current study. Further investigation using simulated data might be needed if the conclusions are intended to be generalized.

Another limitation of the use of real data in the current context is that the examinee true score or proficiency, which is required to evaluate the equity properties, is never known. The error in estimating true scores might have an effect on estimating how well first- or second-order equity is preserved. Future research could investigate this effect by conducting a simulation study.

Range restriction is another limitation. The factors of investigation did not have substantial variability, which might have led to the lack of significant relationships between these factors and the preservation of the three equating properties. The MC-FR correlations ranged from .64 to .94, suggesting moderate to high correlations. The proportion of common items (to MC items) ranged from .23 to .40, and the proportion of MC score points (to composite score points) ranged from .38 to .56. The T_1 statistic, which measured the form difference in this study, ranged from .0139 to .1981. Had a number of tests with more extreme values in these factors (for example, low to modest MC-FR correlations, the proportion of common items being as low as .1 or as high as 1.0, the proportion of MC score points being as low as .1 or as high as .9, or the T_1 value being .5) been included in the analyses, additional significant relationships between these factor and the preservation of the three equating properties might have been observed.

In practice, however, it is difficult to find tests with desired variability in the MC-FR correlation, the proportion of common items, the proportion of MC score points, or form difference. Operationally, if the MC-FR correlation is low, then multidimensionality might be a substantial concern. Also, operationally the proportion of common items is not likely to be in a very wide range. If this proportion is too high, then test security would be an issue, and if it is too

low, then the representativeness of common items would be questionable. Likewise, the proportion of MC items also should be in a reasonable range so that examinees' knowledge and skills are appropriately assessed by each item format. Moreover, operationally it might not be easy to find tests with very large form differences, since test developers attempt to build forms as similar as possible for adequate equating results. Therefore, to address the limitation of range restriction, empirical data with a wide range in the factors of investigation might not be easily available, and a real data study might not be practical. Alternatively, two types of studies could be considered: a simulation study, and a pseudo-test study (e.g., see Sinharay & Holland, 2007).

Moreover, to evaluate the preservation of the first- and second-order equity properties, the true score distribution was estimated assuming IRT models. It is possible that if the criterion is based on a certain psychometric model, it advantages or disadvantages methods that are based on the same model. Therefore, in this study, the two traditional equating methods were compared with each other and the two IRT equating methods were compared with each other, but a fair comparison across the four methods was not able to be performed. Future research could incorporate a strong true score model such as the beta 4 model (Lord, 1965) in addition to the IRT models to estimate the true score distribution, and the D_1 and D_2 indices could be calculated under each model and conclusions could be compared across models. Future research could also develop a procedure that could fairly compare across the traditional and IRT equating methods.

Although unidimensionality held reasonably well given the criteria used for most of the tests used in the study, multidimensionality was found to be present in several tests, which might influence the IRT calibration results, and in turn affect the equating results. It is important to note that conclusions from these tests are limited. Future research could fit multidimensional IRT models to data and use multidimensional IRT equating methods.

References

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

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

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12,* 383-408.

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, The 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.

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

Divgi, D. R. (1981, April). *Two direct procedures for scaling and equating tests with item response theory.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

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.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Harris, D. J. (1991). A comparison of Angoff's Design I and Design I1 for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28,* 221-235.

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

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

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

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. (2004a). POLYCSEM [Computer program]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Kolen, M. J. (2004). POLYEQUATE [Computer Software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285-307.

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

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

Lee, W., Hagge, S. L., He, Y., Kolen, M. J., & Wang, W. (2010, April). *Equating mixed-format tests using dichotomous anchor items*. Paper presented in the structured poster session Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple-Choice and Constructed-Response Items at the Annual Meeting of the American Educational Research Association, Denver, CO.

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

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

McDonald, P. R., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27*, 82-99.

Morris, C. N. (1982). On the foundations of test equating. In P.W. Holland and D.B. Rubin (Eds.), *Testing equating* (pp. 169-191). New York: Academic.

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

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

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

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

Sinharay, S., & Holland, P. W. (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*(3), 249-275.

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

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

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329-346.

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

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 of Measurement in Education, San Diego, CA.

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

Wherry, R. J., & Gaylord, R. H. (1944). Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika*, *9*, 237-244.

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.

Yen, W. M. (1983). Tau-equivalence and equipercentile equating. *Psychometrika,* 48, 353-369.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299-325.

Zhang, J. (2007). *Dichotomous or polytomous model*? *Equating of testlet-based tests in light of conditional item pair correlations*. Unpublished doctoral dissertation, University of Iowa.

Table 1

Number of Items, Maximum FR Score Points, Section Weights, and Total Score Points

Test	Form	# of MC	# of FR	Max FR Score Points	MC Weights	FR Weights	MC Total	FR Total	Composite Total
U.S. History	2004	78	3	9 each	1	4, 2, 2	78	72	150
	2007	80	3	9 each	1	4, 2, 2	80	72	152
Art History	2005	115	9	4, 4, 4, 4, 4, 4, 4, 9, 9	1	4	115	184	299
	2007	115	9	9, 4, 4, 4, 4, 4, 4, 4, 9	1	4	115	184	299
Biology	2004	99	4	10 each	1	2	99	80	179
	2006	98	4	10 each	1	2	98	80	178
	2005	98	4	10 each	1	2	98	80	178
	2007	99	4	10 each	1	2	99	80	179
Chemistry	2004	75	6	10, 10, 15, 9, 10, 8	2	3	150	186	336
	2006	75	6	9, 9, 15, 9, 9, 8	2	3	150	177	327
	2005	75	6	10, 9, 15, 9, 9, 8	2	3	150	180	330
	2007	75	6	9, 10, 10, 15, 9, 9	2	3	150	186	336
English Language & Composition	2004	53	3	9 each	2	5	106	135	241
	2007	52	3	9 each	2	5	104	135	239
Environmental Science	2004	99	4	10 each	1	2	99	80	179
	2006	100	4	10 each	1	2	100	80	180
	2004	99	4	10 each	1	2	99	80	179
	2007	100	4	10 each	1	2	100	80	180
	2005	98	4	10 each	1	2	98	80	178
	2007	100	4	10 each	1	2	100	80	180
European History	2005	80	3	9 each	1	4, 2, 2	80	72	152
	2007	80	3	9 each	1	4, 2, 2	80	72	152
French Language	2004	80	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 1, 5, 3, 3, 3, 3, 3	160	150	310
	2006	84	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 1, 5, 3, 3, 3, 3, 3	168	150	318
	2005	79	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 1, 5, 3, 3, 3, 3, 3	158	150	308
	2007	85	8	15, 15, 9, 5, 5, 5, 5, 5	2	1, 1, 5, 3, 3, 3, 3, 3	170	150	320
German Language	2004	70	9	20, 9, 6, 6, 6, 6, 6, 6, 6	2, 3	1, 10, 1, 1, 1, 1, 1, 1, 8	172	194	366
	2006	70	9	20, 9, 6, 6, 6, 6, 6, 6, 6	2, 3	1, 10, 1, 1, 1, 1, 1, 1, 8	170	194	364
	2005	70	9	20, 9, 6, 6, 6, 6, 6, 6, 6	2, 3	1, 10, 1, 1, 1, 1, 1, 1, 8	170	194	364
	2007	70	9	20, 9, 6, 6, 6, 6, 6, 6, 6	2, 3	1, 10, 1, 1, 1, 1, 1, 1, 8	170	194	364
Comparative Government & Politics	2006	55	8	1, 2, 3, 2, 2, 4, 6, 8	1	2	55	56	111
	2007	55	8	3, 2, 2, 3, 3, 6, 6, 5	1	2	55	60	115
Spanish Language	2004	90	8	20, 9, 9, 4, 4, 4, 4, 4	1	1, 4, 2, 1, 1, 1, 1, 1	90	94	184
	2006	75	8	20, 9, 9, 4, 4, 4, 4, 4	1	1, 4, 2, 1, 1, 1, 1, 1	75	94	169
Spanish Literature	2004	65	7	9, 5, 9, 5, 5, 5, 5	2	5, 4, 5, 4, 5, 5, 4	130	200	330
	2005	65	7	9, 5, 9, 5, 5, 5, 5	2	5, 4, 5, 4, 5, 5, 4	130	200	330
	2004	65	7	9, 5, 9, 5, 5, 5, 5	2	5, 4, 5, 4, 5, 5, 4	130	200	330
	2006	64	7	9, 5, 9, 5, 5, 5, 5	2	5, 4, 5, 4, 5, 5, 4	128	200	328
	2005	65	7	9, 5, 9, 5, 5, 5, 5	2	5, 4, 5, 4, 5, 5, 4	130	200	330
	2007	65	6	9, 5, 9, 5, 9, 5	2	5, 4, 5, 4, 5, 4	130	195	325
World History	2004	70	3	9 each	1	2	70	54	124
	2005	70	3	9 each	1	2	70	54	124
	2004	70	3	9 each	1	2	70	54	124
	2006	70	3	9 each	1	2	70	54	124

Note. For the German Language exam, a weight of 2 was assigned to the first 38 MC items for Form 2004 and to the first 40 MC items for Forms 2005 through 2007, and a weight of 3 was assigned to the latter 32 MC items for Form 2004 and to the latter 30 MC items for Forms 2005 through 2007.

Table 2

Sample Sizes and Descriptive Statistics

				Composite Score		Common-item Score			
Test	Form	N[a]	N[b]	Mean	S.D.	Max Points	Mean	S.D.	Effect Size
U.S. History	2004	20000	1674	62.9659	22.8794	21	10.9892	4.2566	.2226
	2007	20000	1760	69.6455	24.9144	21	11.9744	4.5880	
Art History	2005	15902	3935	146.6770	58.7087	30	17.0003	5.5753	.0710
	2007	17902	3973	131.8663	49.2227	30	16.6096	5.4288	
Biology	2004	20000	2739	85.9704	38.5602	26	14.7174	6.0649	.2003
	2006	20000	3517	93.1271	39.9725	26	15.9340	6.0823	
	2005	20000	2875	85.9990	37.8151	23	13.4717	5.1214	.0834
	2007	20000	3597	82.3083	37.5210	23	13.8966	5.0658	
Chemistry	2004	20000	1767	147.7991	91.1934	50	24.8500	12.6441	.0964
	2006	20000	1139	135.5988	93.9524	50	23.6084	13.1049	
	2005	20000	1186	135.8963	90.9505	50	23.5734	13.0222	.3037
	2007	20000	2117	153.7326	95.5803	50	27.5579	13.2127	
English Language & Composition	2004	20000	6304	149.9254	36.6109	36	27.6206	6.5987	.0415
	2007	20000	6709	142.5676	36.8847	36	27.3451	6.6707	
Environmental Science	2004	20000	3809	74.6051	30.6592	25	13.3392	5.0478	.0826
	2006	20000	4353	85.6492	32.5606	25	13.7641	5.2391	
	2004	20000	3809	74.6051	30.6592	25	13.0449	4.8051	.0872
	2007	20000	3800	84.4234	34.5219	25	13.4782	5.1284	
	2005	20000	4008	77.6317	30.5999	30	16.1305	6.1382	.0579
	2007	20000	3800	84.4234	34.5219	30	16.4913	6.3156	
European History	2005	20000	1638	80.1056	27.1575	20	11.9420	4.1859	.1643
	2007	20000	1107	71.2981	27.8730	20	11.2511	4.2222	
French Language	2004	16750	6474	198.3979	61.8420	48	33.6426	11.1841	.1224
	2006	18595	6180	199.2654	64.3748	48	32.2770	11.1213	
	2005	18068	7619	196.2192	57.3001	52	34.2108	10.9886	.0749
	2007	18878	6477	202.0571	63.5130	52	33.3747	11.3412	
German Language	2004	3899	1744	273.9289	76.4178	54	40.7787	13.4604	.0521
	2006	4450	2097	257.3591	77.9793	54	40.0787	13.4098	
	2005	4048	1994	247.5461	78.9026	51	35.5617	11.8789	.0104
	2007	4078	2059	247.2084	78.3768	51	35.6848	11.7182	
Comparative Government & Politics	2006	12247	3991	58.8449	20.2817	20	12.2130	3.8279	.0384
	2007	12149	4079	57.9684	21.1267	20	12.3592	3.7936	
Spanish Language	2004	20000	9943	124.1802	30.1444	26	17.2124	4.8973	.1075
	2006	20000	8229	114.8769	28.0955	26	16.6814	4.9846	
Spanish Literature	2004	11797	6357	174.9818	44.9919	46	27.8053	8.5405	.0150
	2005	13111	6327	195.0779	43.6756	46	27.9327	8.4606	
	2004	11797	6357	174.9818	44.9919	36	21.9144	5.6346	.0197
	2006	13609	6735	186.4261	44.7219	36	21.8016	5.8116	
	2005	13111	6327	195.0779	43.6756	52	32.9211	8.6918	.0358
	2007	14338	6708	184.0789	46.5419	52	33.2326	8.7040	
World History	2004	20000	5952	60.9824	21.4296	22	14.1489	4.2375	.0162
	2005	20000	6136	62.2994	22.6436	22	14.0794	4.3154	
	2004	20000	5952	60.9824	21.4296	20	12.3290	3.7193	.0208
	2006	20000	5668	62.0473	23.0339	20	12.2495	3.9250	

[a] Original Sample sizes in the AP data.

[b] Actual sample sizes used in the study.

Table 3

Pearson Correlations between MC Scores and FR Scores, Proportions of Common Items to MC Items, and Proportions of MC Score Points to Composite Score Points

Equating Linkages	MC-FR Correlation	% Common (to MC)	% MC (to Composite)
U.S History 04-07	.7743	.2659	.5232
Art History 05-07	.8407	.2609	.3846
Biology 04-06	.8983	.2640	.5518
Biology 05-07	.8838	.2335	.5518
Chemistry 04-06	.9424	.3333	.4526
Chemistry 05-07	.9415	.3333	.4505
English Language & Composition 04-07	.6444	.3429	.4375
Environmental Science 04-06	.8391	.2513	.5543
Environmental Science 04-07	.8457	.2513	.5543
Environmental Science 05-07	.8432	.3031	.5531
European History 05-07	.7350	.2500	.5263
French Language 04-06	.8597	.2929	.5222
French Language 05-07	.8509	.3175	.5221
German Language 04-06	.8550	.3158	.4685
German Language 05-07	.8590	.3000	.4670
Comparative Government & Politics 06-07	.8286	.3636	.4869
Spanish Language 04-06	.7835	.3178	.4665
Spanish Literature 04-05	.6521	.3538	.3939
Spanish Literature 04-06	.6657	.2791	.3921
Spanish Literature 05-07	.6474	.4000	.3970
World History 04-05	.7668	.3143	.5645
World History 04-06	.7692	.2857	.5645

Table 4

Kolmogorov-Smirnov T Statistics before and after Equating

Equating Linkages	Before Equating (T_1)	After Equating (T_2)							
		UnSm FE	UnSm CE	S=.1 FE	S=.1 CE	S=.3 FE	S=.3 CE	IRT True	IRT Obs
U.S History 04-07	.0428	.0419	.0422	.0402	.0353	.0374	.0312	.0135	.0124
Art History 05-07	.0951	.0272	.0209	.0227	.0162	.0216	.0150	.0084	.0067
Biology 04-06	.0139	.0369	.0295	.0302	.0225	.0301	.0222	.0100	.0073
Biology 05-07	.0720	.0240	.0251	.0240	.0237	.0239	.0237	.0092	.0089
Chemistry 04-06	.0206	.0581	.0416	.0554	.0416	.0464	.0352	.0179	.0103
Chemistry 05-07	.0570	.0285	.0384	.0264	.0345	.0284	.0335	.0184	.0056
English Language & Composition 04-07	.0695	.0157	.0164	.0148	.0153	.0114	.0130	.0115	.0086
Environmental Science 04-06	.1040	.0252	.0177	.0238	.0177	.0199	.0162	.0118	.0103
Environmental Science 04-07	.0758	.0242	.0313	.0251	.0266	.0215	.0233	.0117	.0106
Environmental Science 05-07	.0547	.0279	.0293	.0271	.0271	.0236	.0236	.0112	.0108
European History 05-07	.0553	.0578	.0473	.0497	.0400	.0475	.0400	.0148	.0127
French Language 04-06	.0816	.0278	.0228	.0278	.0227	.0270	.0220	.0062	.0061
French Language 05-07	.0821	.0201	.0206	.0192	.0190	.0171	.0161	.0067	.0061
German Language 04-06	.0639	.0278	.0233	.0237	.0187	.0228	.0175	.0139	.0110
German Language 05-07	.0218	.0352	.0352	.0324	.0304	.0275	.0257	.0084	.0053
Comparative Government & Politics 06-07	.0438	.0296	.0305	.0289	.0296	.0247	.0247	.0201	.0182
Spanish Language 04-06	.1206	.0584	.0475	.0575	.0485	.0559	.0451	.0181	.0150
Spanish Literature 04-05	.1981	.0276	.0298	.0266	.0276	.0236	.0236	.0122	.0084
Spanish Literature 04-06	.1442	.0341	.0324	.0339	.0292	.0347	.0298	.0099	.0082
Spanish Literature 05-07	.1155	.0361	.0355	.0361	.0305	.0336	.0305	.0096	.0093
World History 04-05	.0493	.0265	.0294	.0286	.0265	.0244	.0222	.0185	.0155
World History 04-06	.0478	.0278	.0300	.0272	.0257	.0272	.0248	.0157	.0157

Table 5

D_1 for Various Equating Methods

Equating Linkages	First-Order Equity (D_1)							
	UnSm FE	UnSm CE	S=.1 FE	S=.1 CE	S=.3 FE	S=.3 CE	IRT True	IRT Obs
U.S History 04-07	.1134	.0841	.1159	.0852	.1170	.0943	.0055	.0108
Art History 05-07	.0297	.0211	.0293	.0210	.0278	.0185	.0043	.0048
Biology 04-06	.0453	.0298	.0458	.0302	.0435	.0278	.0016	.0063
Biology 05-07	.0408	.0382	.0411	.0384	.0407	.0375	.0022	.0034
Chemistry 04-06	.0443	.0348	.0444	.0346	.0423	.0341	.0026	.0044
Chemistry 05-07	.0494	.0553	.0505	.0555	.0490	.0549	.0062	.0079
English Language & Composition 04-07	.0158	.0143	.0157	.0130	.0172	.0129	.0068	.0116
Environmental Science 04-06	.0278	.0199	.0274	.0173	.0276	.0163	.0046	.0078
Environmental Science 04-07	.0396	.0386	.0397	.0378	.0400	.0383	.0034	.0051
Environmental Science 05-07	.0401	.0339	.0386	.0332	.0380	.0317	.0022	.0059
European History 05-07	.0775	.0631	.0751	.0625	.0755	.0619	.0091	.0086
French Language 04-06	.0446	.0377	.0447	.0379	.0451	.0386	.0023	.0033
French Language 05-07	.0363	.0359	.0355	.0350	.0345	.0338	.0017	.0030
German Language 04-06	.0296	.0360	.0294	.0361	.0264	.0322	.0053	.0038
German Language 05-07	.0553	.0511	.0550	.0504	.0537	.0486	.0019	.0034
Comparative Government & Politics 06-07	.0212	.0144	.0212	.0145	.0186	.0122	.0041	.0105
Spanish Language 04-06	.1446	.1247	.1445	.1247	.1458	.1261	.0085	.0225
Spanish Literature 04-05	.0439	.0550	.0442	.0548	.0419	.0509	.0054	.0100
Spanish Literature 04-06	.0680	.0602	.0681	.0604	.0704	.0622	.0083	.0109
Spanish Literature 05-07	.0508	.0480	.0463	.0427	.0451	.0416	.0054	.0070
World History 04-05	.0204	.0189	.0204	.0191	.0190	.0188	.0054	.0053
World History 04-06	.0327	.0271	.0327	.0272	.0330	.0286	.0073	.0090
Median	.0423	.0368	.0426	.0370	.0413	.0358	.0050	.0066

Table 6

D_2 for Various Equating Methods

Equating Linkages	Second-Order Equity (D_2)							
	UnSm FE	UnSm CE	S=.1 FE	S=.1 CE	S=.3 FE	S=.3 CE	IRT True	IRT Obs
U.S History 04-07	.0164	.0199	.0097	.0124	.0051	.0082	.0337	.0314
Art History 05-07	.0237	.0226	.0238	.0219	.0226	.0199	.0141	.0135
Biology 04-06	.0219	.0211	.0217	.0210	.0207	.0203	.0217	.0200
Biology 05-07	.0240	.0246	.0236	.0244	.0224	.0229	.0108	.0097
Chemistry 04-06	.0225	.0179	.0229	.0184	.0208	.0171	.0129	.0120
Chemistry 05-07	.0156	.0179	.0148	.0166	.0125	.0139	.0133	.0097
English Language & Composition 04-07	.0158	.0195	.0160	.0187	.0148	.0173	.0194	.0130
Environmental Science 04-06	.0240	.0232	.0223	.0216	.0212	.0193	.0182	.0159
Environmental Science 04-07	.0151	.0158	.0133	.0152	.0115	.0141	.0144	.0127
Environmental Science 05-07	.0189	.0197	.0182	.0194	.0170	.0183	.0208	.0190
European History 05-07	.0382	.0388	.0375	.0377	.0368	.0363	.0214	.0171
French Language 04-06	.0117	.0119	.0111	.0118	.0107	.0109	.0127	.0123
French Language 05-07	.0138	.0121	.0130	.0112	.0110	.0089	.0094	.0089
German Language 04-06	.0215	.0236	.0190	.0202	.0149	.0157	.0079	.0056
German Language 05-07	.0148	.0156	.0151	.0150	.0115	.0104	.0144	.0139
Comparative Government & Politics 06-07	.0353	.0326	.0353	.0326	.0341	.0315	.0261	.0226
Spanish Language 04-06	.0338	.0373	.0335	.0369	.0329	.0359	.0678	.0616
Spanish Literature 04-05	.0280	.0284	.0278	.0277	.0264	.0252	.0259	.0223
Spanish Literature 04-06	.0087	.0142	.0085	.0138	.0054	.0120	.0211	.0189
Spanish Literature 05-07	.0337	.0340	.0294	.0294	.0281	.0282	.0139	.0116
World History 04-05	.0198	.0180	.0195	.0180	.0183	.0175	.0117	.0108
World History 04-06	.0107	.0127	.0098	.0117	.0085	.0096	.0194	.0164
Median	.0207	.0198	.0193	.0190	.0177	.0174	.0163	.0137

Table 7

Spearman Correlation Coefficients of D_1, D_2, and T_2 with Various Factors Using 19 Equating Linkages

	Equating	MC-FR Corr		% Common		% MC		T_1	
	Method	*r*	*p* -val	*r*	*p* -val	*r*	*p* -val	*r*	*p* -val
D_1	UnSm FE	-.04	.8866	-.14	.5542	-.13	.6037	.02	.9375
	UnSm CE	-.03	.9092	-.11	.6497	-.16	.5043	.17	.4953
	S=.1 FE	.00	.9943	-.14	.5762	-.11	.6548	.00	.9886
	S=.1 CE	-.03	.9092	-.11	.6497	-.16	.5043	.17	.4953
	S=.3 FE	-.02	.9432	-.16	.5137	-.11	.6419	.04	.8810
	S=.3 CE	-.04	.8585	-.07	.7670	-.21	.3990	.18	.4636
	IRT True	-.68	.0014	.07	.7779	-.25	.2948	.35	.1408
	IRT Obs	-.68	.0014	.15	.5518	-.14	.5542	.13	.6013
D_2	UnSm FE	-.09	.7210	-.03	.9148	.03	.8978	.04	.8810
	UnSm CE	-.26	.2864	-.07	.7724	.01	.9829	.14	.5666
	S=.1 FE	-.04	.8866	.06	.8166	-.07	.7670	.06	.8028
	S=.1 CE	-.17	.4770	-.05	.8444	-.07	.7724	.17	.4953
	S=.3 FE	-.05	.8250	.04	.8556	-.06	.8000	.12	.6318
	S=.3 CE	-.15	.5280	-.02	.9488	-.05	.8444	.18	.4591
	IRT True	-.49	.0320	-.08	.7506	.12	.6291	-.16	.4999
	IRT Obs	-.37	.1226	-.17	.4905	.23	.3494	-.19	.4243
T_2	UnSm FE	.00	.9943	.11	.6679	-.11	.6419	-.31	.2038
	UnSm CE	-.06	.8083	.14	.5787	-.14	.5762	-.25	.3108
	S=.1 FE	-.10	.6785	.05	.8389	-.05	.8305	-.20	.4034
	S=.1 CE	.00	.9886	.22	.3609	-.10	.6836	-.22	.3670
	S=.3 FE	-.05	.8417	.00	.9972	-.12	.6367	-.17	.4770
	S=.3 CE	-.04	.8585	.12	.6138	-.09	.7049	-.16	.5045
	IRT True	-.09	.7157	.31	.2036	.06	.8000	-.29	.2353
	IRT Obs	-.45	.0548	.03	.9007	.37	.1195	-.11	.6576

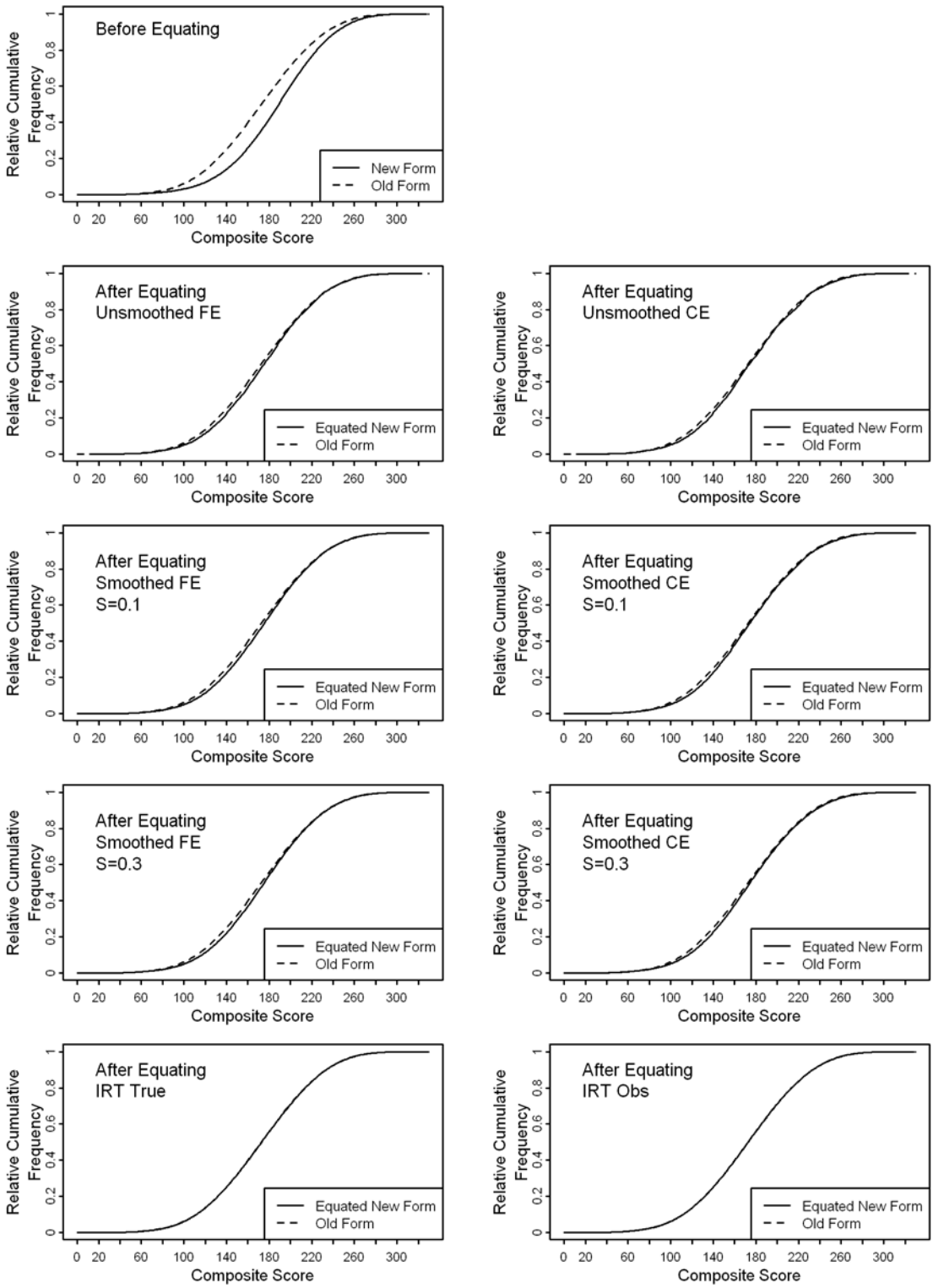

Figure 1. Score distributions of the old and new forms for Spanish Literature 2004-2006 before equating and after equating using various equating methods.

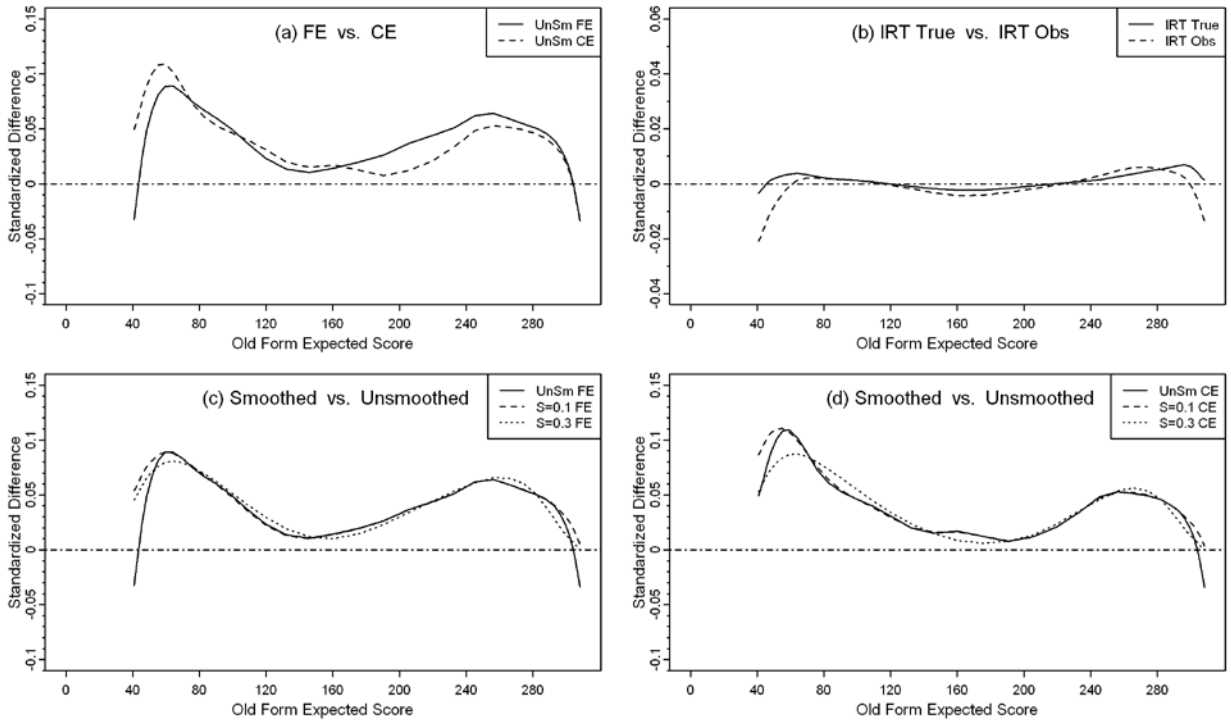

Figure 2. Standardized differences between the old and new forms on the expected scores for French Language 2004-2006.

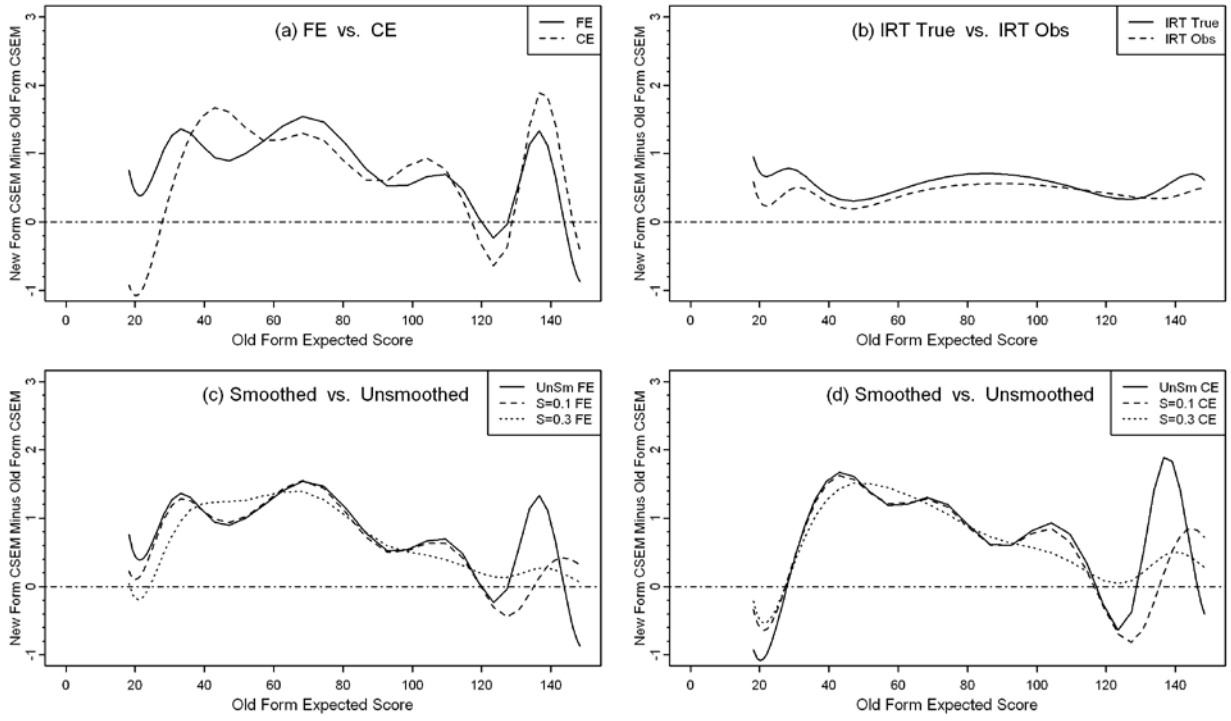

Figure 3. Conditional standard errors of measurement (SEMs) for European History 2005-2007.

Chapter 8: Evaluating Smoothing in Equipercentile Equating Using Fixed Smoothing Parameters

Chunyan Liu and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

This paper compared polynomial loglinear presmoothing and cubic spline postsmoothing using fixed smoothing parameters for equipercentile equating under the random groups design. A wide range of sample sizes (N=100, 200, 500, 1000, 2,000, 5,000, and 10,000) and degrees were investigated. A pseudo-test design based on data from the College Board Advanced Placement Program® (AP®) test forms was used in this study to create criterion equatings based on large sample operational data. The results indicated that both loglinear presmoothing and cubic spline postsmoothing reduced random equating error (average squared error) across all pseudo-tests and all sample sizes with more random error being reduced by more smoothing. With smaller sample sizes, overall equating error (average mean-squared error) tended to be smaller when more smoothing was used for both presmoothing and postsmoothing. Smoothing tended not to work as well in reducing overall equating error for samples with larger sample sizes ($N \geq 5{,}000$), especially when using more smoothing, and in some cases smoothing was found to introduce error compared to no smoothing. Cubic spline postsmoothing tended to perform slightly better than loglinear presmoothing in terms of the reduction of average equating error.

Evaluating Smoothing in Equipercentile Equating Using Fixed Smoothing Parameters

Two sources of error are present, systematic error and random error, when conducting equating. Systematic error can be introduced in several ways, such as smoothing, violation of statistical assumptions, and an improperly implemented data collection design. Random error, often indexed by the standard error of equating (*SEE*), arises whenever examinees are considered a sample from the population.

Smoothing is designed to reduce random error without introducing very much systematic error. Two types of smoothing techniques have been investigated in previous literature, presmoothing and postsmoothing. In presmoothing, the observed score distributions are smoothed prior to conducting equating. The most commonly used presmoothing method is the polynomial loglinear presmoothing (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Darroch & Ratcliff, 1972; Haberman, 1974a, 1974b, 1978; Hanson, 1990; Hanson, Zeng, & Colton, 1994; Holland & Thayer, 1987, 2000; Moses, 2009; Moses & Holland, 2009a, 2009b, 2009c, 2010; Rosenbaum & Thayer, 1987). In postsmoothing, the unsmoothed raw-score equating results are smoothed directly. The most commonly used postsmoothing method is the cubic spline postsmoothing method (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Kolen, 1984; Kolen & Jarjoura, 1987; Hanson et al., 1994; Zeng, 1995). The amount of smoothing for both of the methods is controlled by a parameter that is under the control of the individual conducting equating. Both polynomial loglinear presmoothing and cubic spline postsmoothing have been used in practice and have been shown to have the potential to improve equating accuracy, and the results provided by presmoothing and postsmoothing were of comparable precision (Colton, 1995; Cui & Kolen, 2006; Hanson et al., 1994). However, as emphasized by Kolen and Brennan (2004, p.100), no empirical research-based procedure exists for choosing whether to use presmoothing or postsmoothing. Therefore, more research is necessary to provide more guidance about when to use each of these smoothing methods.

A pseudo-test design based on data from the College Board Advanced Placement Program® (AP®) test forms is used in this study to create criterion equatings based on large sample operational data. The performance of loglinear presmoothing and cubic spline postsmoothing with a broad range of sample sizes and smoothing parameter values for

equipercentile equating under the random groups design are evaluated using the criterion equatings.

Smoothing Methods

Polynomial Loglinear Presmoothing

The polynomial loglinear method was described by Darroch and Ratcliff (1972), Haberman (1974a, 1974b, 1978), Rosenbaum and Thayer (1987), and Holland and Thayer (1987, 2000). In this method, the logarithm of the sample score frequency is fit by a polynomial function. That is, for a particular score point *x*,

$$\log[N\hat{f}(x)] = \omega_0 + \omega_1 x + \omega_2 x^2 + ... + \omega_C x^C , \quad (1)$$

where N is the sample size, $\hat{f}(x)$ is the model implied relative frequency distribution, C is the degree of polynomial typically chosen from 2 to 10, and ω_0, ω_1, …, and ω_C are the estimated parameters for the polynomial function. The Newton-Raphson algorithm is typically used to compute the maximum likelihood estimates (*MLE*) of ω_0, ω_1, …, and ω_C (Haberman, 1974b; Holland & Thayer, 1987).

One desirable feature of the *MLE* of the parameters in Equation 1 is that the first C moments of the resulting fitted distribution are the same as those of the observed score distribution (Darroch & Ratcliff, 1972). For example, if C=4, then the mean, standard deviation, skewness, and kurtosis are identical for the observed score distribution and the fitted score distribution.

Cubic Spline Postsmoothing

Cubic spline smoothing was first described by Reinsch (1967), and the spline fitting algorithm was presented by de Boor (1978, pp. 235-243). This procedure was first applied in the field of equating by Kolen (1984). This method has been investigated in several equating studies (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Hanson et al., 1994; Kolen & Jarjoura, 1987; Zeng, 1995). This smoothing method is summarized as follows, and detailed information can be found in Kolen and Brennan (2004, pp. 84-91).

Given integer scores $x_0, x_1, ..., x_K$ and $x_0 < x_1 < ... < x_K$ where K is the maximum score point, the cubic spline function defined for the interval $x_i < x < x_{i+1}$ is

$$\hat{d}_Y(x) = v_{0i} + v_{1i}(x - x_i) + v_{2i}(x - x_i)^2 + v_{3i}(x - x_i)^3 , \quad (2)$$

where the coefficients v_{0i}, v_{1i}, v_{2i}, and v_{3i} are constants and allowed to change from one interval to the next. At each score point, x_i, the second derivative of the cubic spline function is continuous. Different cubic spline functions are estimated for each interval of $[x_i, x_i+1]$ over the range of score points x_{low} to x_{high}, where x_{low} is the lowest score in the range, x_{high} is the highest score point in the range, and $0 \le x_{low} \le x \le x_{high} \le K$.

The cubic spline function is obtained by requiring minimum curvature and satisfying the following inequality:

$$\frac{\sum_{i=low}^{high}\left[\frac{\hat{d}_Y(x_i)-\hat{e}_Y(x_i)}{\widehat{se}\left[\hat{e}_Y(x_i)\right]}\right]^2}{x_{high}-x_{low}+1} \le S. \qquad (3)$$

The summation is over the range of scores that is used to fit the spline function. In the context of equipercentile equating, $\hat{e}_Y(x_i)$ is the equated raw score of x_i on Form X; $\hat{d}_Y(x_i)$ is the equated score of x_i after smoothing; and $\widehat{se}\left[\hat{e}_Y(x_i)\right]$ is the estimated standard error of the equated raw score.

Parameter S ($S \ge 0$) is the smoothing parameter for cubic spline postsmoothing, which controls the degree of smoothing. S=0 is no smoothing, where the smoothed equivalents equal the unsmoothed equivalents. If S is very large, the spline function is a straight line. Several studies (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Hanson et al., 1994; Kolen, 1991; Kolen & Jarjoura, 1987; Kolen & Brennan, 2004; Zeng, 1995) suggested that a value of S between 0 and 1 yields appropriate results in equating. The spline function is typically fit over the score points with percentile ranks between .5 and 99.5 (Kolen, 1984). A linear interpolation procedure is often used to obtain the smoothed equated score outside of this range.

Method

Data

Data for this study were from the College Board Advanced Placement Program® (AP®) test forms. The AP tests considered are mixed-format tests containing two types of items: multiple-choice items (MC) and free-response (FR) items. Two forms of each of two AP tests were used: AP Biology and AP Environmental Science. For each test, the test forms

administered in 2005 and 2006 were used. In all four forms, the first 100 items were MC items, and the last 4 items were FR items. Each MC item was scored either 0 or 1, and each FR item was scored from 0 to 10. Sample sizes are shown in Table 1.

Formula scoring was used for operational scoring of the MC items for these test forms. In the directions for the examinations, examinees were instructed that they could guess on the MC items but would be penalized for incorrect guessing. For the tests used in this chapter, during operational scoring, .25 points were subtracted from examinees' scores if their response to each MC item was wrong. If they did not answer the item, there was no penalty. Consequently, there were a large number of missing responses on MC items for some examinees in the operational scoring.

In this study, number correct scoring was used for MC items, which means that correct responses were scored as 1, and incorrect responses were scored as 0. There was no penalty for incorrect responses. Also, only those examinees that responded to at least 80% of the MC items were included. Even for these examinees, there were still quite a few missing responses in their MC responses. Therefore, a two-way imputation procedure (Bernaards & Sijtsma, 2000; Sijtsma & van der Ark, 2003) was used to impute the missing responses for MC items.

The person mean, denoted PM_i, was computed based on all items responded by person i. The item mean, IM_j, was the average score of item j over the examinees who responded to it. The overall mean, OM, was defined as the average of all observed item scores. Then for person i, TW_{ij} was defined for any item j with a missing response as,

$$TW_{ij} = PM_i + IM_j - OM. \quad (4)$$

If TW_{ij} was equal to or greater than .5, a score of 1 (correct) was assigned to the examinee on that item. If TW_{ij} was less than .5, a score of 0 was assigned. The examinees' observed scores were computed based on the item responses after imputation. For simplicity, the total score is the sum of the MC and FR scores.

Construction of Pseudo-Tests

The original test forms were used to create pseudo-test forms. For each original test form, two pseudo-test forms (Form X and Form Y) with the same number of MC and FR items on each form were created; the items on one pseudo-test form were different from the items on the other pseudo-test form created from an original form.

Pseudo-test forms for Biology 2005 and Environmental Science 2005 were constructed to be long test forms with 48 MC items and 2 FR items and are referred to as Form X and Form Y of Biology long (BL) and Environmental Science long (ESL). Pseudo-test forms for Biology 2006 and Environmental Science 2006 were constructed to be short test forms with 24 MC items and 1 FR item and are referred to as Form X and Form Y of Biology short (BS) and Environmental Science short (ESS). Because the FR items were scored from 0 to 10, the maximum score is 68 for tests BL and ESL and 34 for BS and ESS. Therefore, the long forms are twice as long as the short forms in terms of number of MC items, number of FR items, and the maximum score. Table 1 describes the original and pseudo-test forms.

The pseudo-test forms for Biology 2005 and Environmental Science 2005 were based on an even-odd item number split. Form X for BL and ESL contained MC items numbered 1, 3, 5, …, 95, and the first and third FR items in the original test. Form Y for BL and ESL contained MC items numbered 2, 4, 6, …, 96, and the second and fourth FR items in the original test. There were 48 MC items and 2 FR items in the long forms. For Biology 2006 and Environmental Science 2006, the pseudo-test forms were created in such a way that items 1, 5, 9,…,89, 93, and the first FR item were in Form X of BS and ESS; items 3, 7, 11,…,91, 95, and the second FR item were in Form Y of BS and ESS. There were 24 MC items and 1 FR items in the short forms.

Therefore, based on the original tests, four pseudo tests were created: two long pseudo-tests (BL and ESL) and two short pseudo-tests (BS and ESS). For each pseudo-test, there were two test forms, Form X and Form Y. Each examinee who took an original test has scores on both Form X and Form Y. Distributions of these scores, which were based on large examinee samples taking both Form X and Form Y, were used as population distributions and to define the criterion equating using unsmoothed equipercentile procedures. Thus, the criterion equating is a large sample single group equipercentile equating.

Figure 1 shows the distributions of the four pseudo tests for the whole dataset, and Table 2 provides the summary statistics. The effect size (ES) between the two forms was computed as the ratio of difference of means between Form X and Form Y to the pooled standard deviation (Yen, 1986), and the reliability was calculated based on the IRT method (Kolen & Brennan, 2004, pp. 301-302).

Sampling Procedures

As already indicated, the distributions of scores on the pseudo-test forms based on the whole dataset were considered as the population score distributions from which the true equating relationship was developed. The following steps were applied to evaluate the smoothing methods for equipercentile equating under the random groups design for all four pseudo-tests:

1. Random samples of size *N* were drawn from the population distributions separately for Form X and Form Y.
2. Equipercentile equating was conducted using the two sample distributions.
3. Cubic spline postsmoothing with *S*=.01, .05, .1, .2, .3, .5, .75, and 1.0 were applied to the equating results obtained in Step 2.
4. The distributions for the two samples obtained in Step 1 were smoothed using polynomial loglinear presmoothing with *C* = 2, 3, 4, 5, 6, 7, 8, and 9.
5. Equipercentile equating was conducted using the smoothed distributions obtained in Step 4.
6. Steps 1 to 5 were repeated 500 times for each sample size (*N*=100, 200, 500, 1,000, 2,000, 5,000, and 10,000) and each test (BL, BS, ESL, and ESS).

All analyses were conducted using the open source *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009) C code.

Evaluation Criteria

At score point x_i, define the relative frequency of the population distribution for Form X as $f(x_i)$ and the true equating relationship as $e_Y(x_i)$. For the r^{th} pair of samples, define the equating relationship without smoothing as $\hat{e}_{Yr}(x_i)$ and the equating relationship with smoothing as $\hat{t}_{Yr}(x_i)$. The average estimated equipercentile equating relationship with smoothing, $\overline{t}_Y(x_i)$, at score point x_i is defined as:

$$\overline{t}_Y(x_i) = \frac{1}{500}\sum_{r=1}^{500}\hat{t}_{Yr}(x_i), \tag{5}$$

and the *Bias* of the estimated equipercentile equating relationship with smoothing at score point x_i is defined as:

$$Bias[t_Y(x_i)] = \overline{t}_Y(x_i) - e_Y(x_i). \tag{6}$$

The average squared bias (*ASB*) for the estimated equipercentile equating relationship with smoothing is defined as:

$$ASB[t_Y(x)] = \sum_{i=0}^{K} f(x_i) Bias\left[t_Y(x_i)\right]^2, \quad (7)$$

where *K* is the maximum score on the test.

The standard error of equipercentile equating (*SEE*) with smoothing at score point x_i is given by,

$$SEE(x_i) = \sqrt{\frac{\sum_{r=1}^{500}[\hat{t}_{Yr}(x_i) - \overline{t}(x_i)]^2}{500-1}}, \quad (8)$$

and the average squared error (*ASE*) for the estimated equipercentile equating relationship with smoothing is defined in the same way as the average squared bias for equating (Equation 7) , that is,

$$ASE[t_Y(x)] = \sum_{i=0}^{K} f(x_i) SEE(x_i)^2 \ . \quad (9)$$

The root mean square error of equating (*RMSE*) with smoothing at score point x_i is defined by,

$$RMSE[t_Y(x)] = \sqrt{Bias\left[t_Y(x_i)\right]^2 + SEE(x_i)^2} \quad (10)$$

and the average mean squared error (*AMSE*) is defined as the sum of *ASB* and *ASE* for the estimated equipercentile equating relationship with smoothing, that is

$$AMSE[t_Y(x)] = ASB[t_Y(x)] + ASE[t_Y(x)]. \quad (11)$$

These criteria were applicable to both the loglinear presmoothing and cubic spline postsmoothing.

Results

Average Squared Bias (*ASB*)

ASB for the equating results without smoothing (denoted as “none”), with loglinear presmoothing (*C*=2, 3,…, 8, and 9), and with cubic spline postsmoothing (*S*=.01, .05, .1, .2, .3, .5, .75, and 1.0) for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are provided in Tables 3 through 6, respectively. For the results without smoothing, equating bias tended to decrease as sample size increased for all pseudo-test forms.

For polynomial loglinear presmoothing, *ASB* tended to decrease slightly from sample size of 100 to 200 for most values of *C*. However, for sample sizes that are larger than 500, *ASB* tended to be similar across sample sizes for most values of *C*. Compared to no smoothing, loglinear presmoothing did not appear to introduce equating bias for smaller sample sizes (*N*<200). For example, for the BL test, *ASB* was smaller than the results without smoothing for *C*>4 at *N*=100. More smoothing (lower values of *C*) tended to be associated with more equating bias although this was not always the case.

For cubic spline postsmoothing, *ASB* tended to decrease as the sample sizes increased across all pseudo-tests. Similar to the results found for the loglinear presmoothing, the cubic spline postsmoothing not necessarily introduced more equating bias either, especially for smaller sample sizes (*N*<200). For example, for the BL test, *ASB* was smaller than the results without smoothing for all *S* values at *N*=100. More smoothing (larger values of *S*) introduced more equating bias than less smoothing consistently (and more consistently than presmoothing) across all sample sizes and all pseudo-tests.

Average Squared Error (*ASE*)

ASE for the equating results without smoothing (denoted as "none"), with loglinear presmoothing (*C*=2, 3,…, 8, and 9), and with cubic spline postsmoothing (*S*=.01, .05, .1, .2, .3, .5, .75, and 1.0) for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are provided in Tables 7 through 10, respectively. For the results without smoothing, random equating error tended to decrease with the increase of the sample sizes across all pseudo-tests.

For polynomial loglinear presmoothing, *ASE* tended to decrease significantly as sample size increased across all investigated equating conditions. Loglinear presmoothing reduced random equating error consistently. For example, for the BL test, loglinear presmoothing at each *C* value yielded smaller *ASE* than those without smoothing at all sample sizes. In addition, *ASE* tended to increase as the *C* parameter increased (from 2 to 9), which suggests that the more smoothing leads to less random equating error. For example, for the BL test at sample size of 100, the *ASE*'s are 4.1081, 4.5947, 5.1475, 5.3629, 5.6202, 5.7379, 5.8932, 5.9995 for *C*=2, 3,…, 8, 9 for loglinear presmoothing respectively.

Similar results were found for cubic spline postsmoothing: 1) *ASE* tended to decrease significantly as sample size increased across all investigated equating conditions; 2) cubic spline postsmoothing reduced random equating error consistently; 3) *ASE* tended to decrease as the *S*

parameter increased (from .01 to 1) which suggested that more smoothing could reduce more random equating error. For example, the *ASE*'s are 6.0404, 5.5953, 5.1219, 4.4396, 4.0863, 3.8038, 3.7019, and 3.6746 for *S*=.01, .05, .1, .2, .3, .5, .75, and 1.0 respectively for cubic spline postsmoothing with the BL test at sample size of 100. However, no clear conclusion can be made about which smoothing method works better in terms of the reduction of random equating error.

Average Mean Squared Error (*AMSE*)

AMSE for the equating results without smoothing (denoted as "none"), with loglinear presmoothing (*C*=2, 3,…, 8, and 9), and with cubic spline postsmoothing (*S*=.01, .05, .1, .2, .3, .5, .75, and 1.0) for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are given in Tables 11 through 14, respectively. For all results with and without smoothing, *AMSE* tended to decrease with increasing sample sizes across all pseudo-tests because *AMSE* was dominated by the random equating error.

For polynomial loglinear presmoothing, more smoothing (smaller values of *C*) tended to lead to smaller *AMSE* for smaller sample sizes, which is consistent with previous research (Cui & Kolen, 2008). For larger sample sizes, more moderate degrees of smoothing tend to lead to smaller *AMSE*. For example, the smallest *AMSE* was yielded by *C*=2, 3, 3, 3, 4, 4, and 7 for *N*=100, 200, 500, 1,000, 2,000, 5,000, and 10,000, respectively, for the ESS test. Smoothing tends to reduce *AMSE* compared to no smoothing for samples with smaller sample sizes (*N*≤2,000). However, smoothing can have larger *AMSE* than no smoothing for larger sample sizes (*N*≥5,000), especially when more smoothing is used (e.g., *C*≤3). For example, *AMSE* for *C*=2 are .1155 and .1055 for the ESS test at sample sizes of 5,000 and 10,000, respectively, which are larger than those for no smoothing (.0322 for *N*=5,000 and .0163 for *N*=10,000). Therefore, compared to no smoothing, loglinear presmoothing tended to reduce *AMSE* for equating conditions with smaller sample sizes (*N*≤2,000) and larger *C* values (*C*≥4), which is true for all four pseudo-tests investigated in this study.

For cubic spline postsmoothing, similar results were found: 1) More smoothing was associated with smaller *AMSE* for samples with smaller sample sizes, which is also consistent with previous research (Zeng, 1995). For the ESS test, the smallest *AMSE* were yielded by *S*=1.0, .3, .3, .2, .2, .2, and .1 for *N*=100, 200, 500, 1,000, 2,000, 5,000, and 10,000, respectively; 2) smoothing tended to reduce *AMSE* for samples with smaller sample sizes (*N*≤2,000), and led to larger *AMSE* for samples with larger sample sizes (*N*≥5,000), especially for the conditions with

more smoothing ($S \geq .75$). For example, the *AMSE* for $S = .75$ are .0385 and .0202 for the ESS test at sample sizes of 5,000 and 10,000, respectively, which are larger than those without smoothing (.0322 for $N = 5{,}000$ and .0163 for $N = 10{,}000$). 3) cubic spline postsmoothing tended to reduce *AMSE* for equating conditions with smaller sample sizes ($N \leq 2{,}000$) and smaller S values ($S \leq .5$), which is found for all four pseudo-tests investigated in this study.

Since both loglinear presmoothing and cubic spline postsmoothing had smaller *AMSE* for smaller sample sizes ($N \leq 2{,}000$) and less degrees of smoothing ($C \geq 4$ and $S \leq .5$), the performance of loglinear presmoothing with $C \geq 4$ was compared to the performance of cubic spline postsmoothing with $S \leq .5$ for $N \leq 2{,}000$ with the criterion being the *AMSE*. The smallest *AMSE* for these conditions are bolded in Tables 11 through 14. In most situations, cubic spline postsmoothing yielded smaller *AMSE* than loglinear presmoothing. However, for the ESL test, presmoothing with $C = 4$ produced the smallest *AMSE* at $N = 1{,}000$ and 2,000. This finding indicates that cubic spline postsmoothing tends to perform slightly better than loglinear presmoothing in terms of the reduction of *AMSE*.

Discussion and Conclusions

The primary purpose of this study was to compare polynomial loglinear presmoothing and cubic spline postsmoothing using a wide range of sample sizes and a wide range of degrees of smoothing using the fixed smoothing parameter method under the random groups design. It is commonly accepted that smoothing, including both presmoothing and postsmoothing, can reduce total equating error by reducing random equating error without introducing too much systematic error. The results of this study indicate that both loglinear presmoothing and cubic spline postsmoothing reduced systematic error when the sample size was small and smoothing degree was not too large (for example, $C \geq 4$ and $S \leq .2$ at $N = 100$ for the ESL test). The results also indicate that both presmoothing and postsmoothing methods can reduce random equating error (as indexed by *ASE*) across all pseudo-tests and all sample sizes with more random error being reduced by more smoothing which is consistent with previous studies (Cui & Kolen, 2008; Hanson et al., 1994; Kolen & Brennan, 2004; Zeng, 1995). However, no clear conclusion can be made about which smoothing method yields smaller random error.

The *AMSE* of equating suggests that more smoothing is needed for samples with smaller sample sizes for both presmoothing and postsmoothing to reduce overall average equating error, which is consistent with previous studies (Cui & Kolen, 2008; Zeng, 1995). Smoothing tends not

to reduce *AMSE* substantially for samples with larger sample sizes ($N \geq 5{,}000$), especially when $C \leq 3$ and $S \geq .75$, and in some cases can increase *AMSE*. This finding may be due to the observed score distributions and the corresponding equating relationships being smooth enough with large samples that, in some cases, smoothing introduces more bias compared to the reduction of the random error.

The comparison between loglinear presmoothing and cubic spline postsmoothing for smaller sample sizes ($N \leq 2{,}000$) and smaller amount of smoothing ($C \geq 4$ and $S \leq .5$) suggests that cubic spline postsmoothing tends to perform slightly better than the loglinear presmoothing in terms of the reduction of total equating error as indexed by *AMSE*.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Colton, D. A. (1995). *Comparing smoothing and equating methods using small sample sizes*. Unpublished Ph.D. Dissertation. University of Iowa, Iowa City, Iowa.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: the cubic B-spline presmoothing method and the direct presmoothing*. Unpublished doctoral dissertation, University of Iowa, Iowa City, Iowa.

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

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics, 43*, 1470-1480.

De Boor, C. (1978). *A practical guide to splines* (Applied Mathematical Science, Volume 27). New York: Springer-Verlag.

Haberman, S. J. (1974a). *The analysis of frequency data*. Chicago: University of Chicago.

Haberman, S. J. (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30*, 589-600.

Haberman, S. J. (1978). *Analysis of qualitative data. Vol. 1. Introductory topics*. New York: Academic.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions*. (ACT Research Report 90-4). Iowa City, IA: American College Testing.

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

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Technical Report 87-89). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

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

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28,* 257-272.

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

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equatingunder the common item nonequivalent populations design. *Psychometrika, 52,* 43-59.

Moses, T. (2009). A comparison of statistical significance tests for selecting equating functions. *Applied Psychological Measurement, 33,* 285-306.

Moses, T., & Holland, P. W. (2009a). *Selection strategies for bivariate loglinear smoothing models and their effects on NEST equating functions*. (Technical Report 09-04). Princeton, NJ: Educational Testing Service.

Moses, T., & Holland, P. W. (2009b). *Alternative loglinear smoothing models and their effect on equating function accuracy*. (Technical Report 09-48). Princeton, NJ: Educational Testing Service.

Moses, T., & Holland, P. W. (2009c). Selection strategies for Univariate loglinear smoothing models and their effects on equating function accuracy. *Journal of Educational Measurement. 46,* 159-176.

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

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10*, 177-183.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British journal of Mathematical and Statistical Psychology, 40,* 43-49.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325.

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement, 19,* 177-190.

Table 1

Description of Biology and Environmental Science Pseudo-Tests

Original Test	Year	Pseudo-test name	Sample size (*N*)	# MC items	# CR items	Total points
Biology	2005	BL	16,185	48	2	68
	2006	BS	16,899	24	1	34
Environmental Science	2005	ESL	17,610	48	2	68
	2006	ESS	17,508	24	1	34

Table 2

Descriptive Statistics for the Pseudo-Test Forms

Test	*N*	Form	Mean	SD	Skewness	Kurtosis	Reliability[a]	ES
BL	16,185	X	40.4659	11.7178	-.3905	-.3952	.8859	.17
		Y	38.5002	11.5961	-.3248	-.4098	.8919	
BS	16,899	X	20.3859	6.6352	-.4095	-.4821	.8444	.27
		Y	18.7509	5.5931	-.2803	-.3016	.7796	
ESL	17,610	X	36.2289	10.6366	-.2537	-.4245	.8746	.10
		Y	35.1228	11.3824	-.1300	-.6600	.8922	
ESS	17,508	X	17.7731	5.9050	-.0817	-.4976	.7926	-.09
		Y	18.2930	5.9014	-.2700	-.4739	.7966	

[a] The reliabilities were computed using IRT.

Table 3

Average Squared Bias of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	.0645	.0569	.0554	.0644	.0443	.0537	.0526	.0418	.0475	.0357	.0359	.0384	.0382	.0412	.0495	.0572	.0605
200	.0255	.0475	.0381	.0487	.0235	.0306	.0214	.0165	.0203	.0120	.0151	.0191	.0268	.0357	.0528	.0642	.0683
500	.0077	.0412	.0353	.0408	.0174	.0224	.0115	.0080	.0094	.0039	.0064	.0101	.0192	.0288	.0448	.0579	.0643
1,000	.0029	.0380	.0349	.0393	.0151	.0200	.0095	.0064	.0070	.0020	.0043	.0078	.0166	.0252	.0419	.0554	.0620
2,000	.0019	.0392	.0348	.0398	.0149	.0204	.0097	.0069	.0070	.0023	.0044	.0072	.0136	.0198	.0339	.0492	.0586
5,000	.0005	.0389	.0342	.0398	.0148	.0200	.0094	.0065	.0063	.0018	.0035	.0056	.0105	.0152	.0233	.0336	.0437
10,000	.0002	.0391	.0339	.0397	.0147	.0201	.0095	.0066	.0062	.0014	.0027	.0042	.0073	.0106	.0163	.0226	.0289

Table 4

Average Squared Bias of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	.0986	.2854	.1528	.0579	.0632	.0583	.0645	.0586	.0580	.0372	.0391	.0495	.0970	.1561	.2381	.2839	.2985
200	.0258	.2527	.1138	.0188	.0221	.0176	.0197	.0134	.0125	.0032	.0049	.0115	.0479	.0963	.1773	.2287	.2476
500	.0107	.2570	.1066	.0184	.0198	.0164	.0186	.0119	.0111	.0028	.0043	.0084	.0286	.0635	.1405	.2092	.2453
1,000	.0049	.2592	.1122	.0156	.0163	.0159	.0174	.0107	.0099	.0024	.0037	.0068	.0221	.0470	.1038	.1658	.2094
2,000	.0025	.2607	.1117	.0161	.0165	.0161	.0178	.0109	.0099	.0023	.0037	.0067	.0169	.0325	.0703	.1182	.1621
5,000	.0005	.2594	.1080	.0160	.0165	.0153	.0170	.0102	.0090	.0013	.0023	.0041	.0099	.0176	.0356	.0591	.0822
10,000	.0002	.2597	.1084	.0159	.0164	.0151	.0168	.0100	.0088	.0013	.0020	.0031	.0065	.0106	.0201	.0328	.0456

Table 5

Average Squared Bias of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	.0115	.0372	.0383	.0178	.0145	.0072	.0071	.0090	.0096	.0083	.0097	.0126	.0260	.0438	.0738	.0940	.1038
200	.0027	.0385	.0382	.0150	.0124	.0044	.0044	.0049	.0049	.0037	.0046	.0072	.0177	.0318	.0587	.0807	.0931
500	.0013	.0383	.0374	.0159	.0124	.0050	.0050	.0042	.0031	.0036	.0044	.0064	.0126	.0213	.0408	.0627	.0787
1,000	.0008	.0388	.0383	.0151	.0124	.0047	.0049	.0037	.0025	.0031	.0036	.0052	.0097	.0155	.0291	.0465	.0613
2,000	.0002	.0386	.0382	.0149	.0119	.0044	.0046	.0033	.0021	.0028	.0032	.0042	.0070	.0100	.0170	.0267	.0367
5,000	.0001	.0383	.0379	.0151	.0120	.0043	.0044	.0032	.0021	.0031	.0032	.0037	.0052	.0067	.0097	.0136	.0177
10,000	.0000	.0381	.0379	.0151	.0120	.0043	.0043	.0031	.0020	.0031	.0031	.0034	.0042	.0050	.0067	.0088	.0108

Table 6

Average Squared Bias of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	.0145	.0985	.0196	.0099	.0100	.0093	.0101	.0107	.0112	.0088	.0089	.0105	.0230	.0438	.0799	.1016	.1101
200	.0038	.0941	.0117	.0042	.0052	.0029	.0025	.0030	.0026	.0010	.0012	.0024	.0103	.0266	.0606	.0865	.0989
500	.0011	.0945	.0119	.0041	.0049	.0030	.0023	.0027	.0023	.0013	.0017	.0027	.0076	.0174	.0430	.0698	.0864
1,000	.0005	.0949	.0123	.0040	.0047	.0032	.0022	.0023	.0018	.0011	.0015	.0024	.0059	.0127	.0318	.0556	.0744
2,000	.0002	.0946	.0120	.0039	.0047	.0030	.0021	.0020	.0015	.0008	.0012	.0018	.0038	.0072	.0172	.0315	.0456
5,000	.0002	.0951	.0126	.0039	.0046	.0030	.0021	.0020	.0016	.0009	.0012	.0016	.0029	.0046	.0091	.0155	.0220
10,000	.0000	.0948	.0122	.0038	.0045	.0029	.0019	.0018	.0014	.0008	.0010	.0013	.0021	.0029	.0049	.0079	.0111

Table 7

Average Squared Errors of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	6.5296	4.1081	4.5947	5.1475	5.3629	5.6202	5.7379	5.8932	5.9995	6.0404	5.5953	5.1219	4.4396	4.0863	3.8038	3.7019	3.6746
200	3.2520	2.0542	2.2741	2.4952	2.6079	2.7591	2.8217	2.9020	2.9465	3.0197	2.8041	2.5810	2.2462	2.0552	1.9018	1.8474	1.8332
500	1.2015	.7059	.8029	.9003	.9417	.9933	1.0190	1.0475	1.0657	1.1179	1.0424	.9640	.8353	.7515	.6712	.6356	.6221
1,000	.6210	.3856	.4258	.4675	.4886	.5168	.5284	.5442	.5520	.5802	.5452	.5090	.4492	.4104	.3693	.3494	.3425
2,000	.3149	.1930	.2164	.2361	.2468	.2591	.2655	.2739	.2784	.2956	.2790	.2619	.2353	.2157	.1929	.1787	.1720
5,000	.1344	.0860	.0945	.1019	.1070	.1127	.1151	.1177	.1196	.1267	.1215	.1164	.1088	.1030	.0947	.0882	.0836
10,000	.0613	.0368	.0408	.0444	.0471	.0498	.0512	.0528	.0537	.0576	.0555	.0534	.0504	.0481	.0445	.0416	.0398

Table 8

Average Squared Errors of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	5.9137	3.7391	4.1458	4.5631	4.7867	5.0717	5.1830	5.3083	5.3622	5.4226	5.0075	4.5854	4.0381	3.7505	3.4964	3.3859	3.3462
200	2.8793	1.7900	1.9820	2.1781	2.2820	2.4193	2.4727	2.5370	2.5705	2.6653	2.4685	2.2679	2.0004	1.8416	1.7018	1.6396	1.6166
500	1.1602	.7399	.8023	.8747	.9064	.9660	.9878	1.0114	1.0252	1.0778	1.0087	.9404	.8492	.7988	.7408	.6996	.6785
1,000	.5807	.3594	.3967	.4362	.4559	.4848	.4960	.5044	.5114	.5426	.5095	.4757	.4342	.4122	.3855	.3618	.3441
2,000	.2736	.1658	.1822	.2010	.2097	.2231	.2296	.2346	.2386	.2565	.2412	.2250	.2055	.1962	.1858	.1770	.1698
5,000	.1117	.0691	.0770	.0842	.0879	.0925	.0945	.0962	.0976	.1060	.1010	.0961	.0899	.0869	.0837	.0817	.0806
10,000	.0568	.0352	.0389	.0424	.0441	.0465	.0477	.0487	.0494	.0545	.0524	.0503	.0474	.0455	.0437	.0425	.0418

Table 9

Average Squared Errors of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	1.5076	.9795	1.0869	1.2017	1.2485	1.3199	1.3578	1.3882	1.4049	1.3868	1.2779	1.1791	1.0546	.9858	.9214	.8915	.8784
200	.7473	.4664	.5261	.5816	.6082	.6423	.6601	.6761	.6854	.6890	.6365	.5906	.5321	.4992	.4663	.4481	.4396
500	.2942	.1796	.2040	.2257	.2365	.2498	.2581	.2651	.2680	.2732	.2547	.2389	.2196	.2079	.1928	.1820	.1752
1,000	.1449	.0881	.1002	.1109	.1160	.1229	.1260	.1296	.1319	.1348	.1265	.1199	.1123	.1080	.1024	.0970	.0929
2,000	.0758	.0480	.0538	.0588	.0616	.0650	.0666	.0685	.0695	.0715	.0678	.0651	.0621	.0604	.0585	.0567	.0552
5,000	.0276	.0165	.0186	.0205	.0216	.0230	.0237	.0244	.0249	.0261	.0249	.0241	.0231	.0225	.0217	.0212	.0209
10,000	.0150	.0097	.0108	.0116	.0122	.0129	.0133	.0136	.0139	.0145	.0140	.0137	.0134	.0131	.0128	.0125	.0123

Table 10

Average Squared Errors of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	1.5771	1.0257	1.1457	1.2499	1.3038	1.3859	1.4185	1.4593	1.4803	1.4451	1.3265	1.2228	1.0937	1.0264	.9679	.9389	.9286
200	.7735	.4887	.5502	.6021	.6339	.6710	.6860	.7051	.7134	.7107	.6560	.6031	.5377	.5036	.4713	.4530	.4457
500	.2965	.1843	.2090	.2290	.2396	.2534	.2603	.2662	.2705	.2742	.2548	.2359	.2135	.2022	.1883	.1770	.1705
1,000	.1549	.0968	.1095	.1194	.1254	.1321	.1357	.1385	.1407	.1442	.1346	.1259	.1156	.1111	.1061	.1015	.0976
2,000	.0760	.0486	.0546	.0588	.0616	.0650	.0666	.0680	.0690	.0711	.0667	.0627	.0579	.0557	.0538	.0526	.0515
5,000	.0320	.0204	.0230	.0249	.0260	.0273	.0280	.0287	.0291	.0301	.0287	.0273	.0256	.0246	.0236	.0230	.0228
10,000	.0163	.0107	.0120	.0129	.0135	.0141	.0145	.0148	.0150	.0155	.0149	.0144	.0136	.0132	.0126	.0123	.0121

Table 11

Average Mean Squared Errors of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	6.5941	4.1650	4.6501	5.2119	5.4072	5.6739	5.7905	5.9350	6.0470	6.0762	5.6312	5.1603	4.4778	4.1275	**3.8534**	3.7591	3.7351
200	3.2775	2.1017	2.3123	2.5439	2.6314	2.7897	2.8430	2.9185	2.9668	3.0317	2.8192	2.6002	2.2730	2.0909	**1.9546**	1.9116	1.9015
500	1.2092	.7471	.8382	.9412	.9591	1.0157	1.0305	1.0555	1.0750	1.1218	1.0488	.9742	.8545	.7804	**.7160**	.6935	.6863
1,000	.6239	.4236	.4607	.5067	.5036	.5368	.5379	.5505	.5590	.5822	.5495	.5169	.4658	.4356	**.4112**	.4048	.4045
2,000	.3168	.2322	.2512	.2759	.2616	.2795	.2752	.2808	.2854	.2979	.2834	.2691	.2488	.2355	**.2268**	.2279	.2306
5,000	.1349	.1249	.1286	.1417	.1218	.1327	.1245	.1243	.1259	.1285	.1250	.1220	.1193	.1182	.1180	.1218	.1273
10,000	.0615	.0759	.0747	.0841	.0618	.0699	.0606	.0593	.0599	.0590	.0582	.0576	.0577	.0587	.0608	.0641	.0688

Note. Bold font indicates the smallest value for $C \geq 4$ or $S \leq .5$ at a given sample size less than 2,000; Number in shaded cell means that smoothing increases total equating error at a certain sample size and a certain smoothing parameter.

Table 12

Average Mean Squared Errors of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	6.0123	4.0245	4.2986	4.6210	4.8499	5.1300	5.2475	5.3669	5.4202	5.4598	5.0466	4.6349	4.1351	3.9066	**3.7345**	3.6697	3.6447
200	2.9051	2.0426	2.0958	2.1969	2.3041	2.4369	2.4924	2.5504	2.5830	2.6685	2.4734	2.2794	2.0483	1.9378	**1.8791**	1.8683	1.8642
500	1.1709	.9969	.9089	.8931	.9262	.9824	1.0064	1.0234	1.0363	1.0806	1.0130	.9488	.8778	**.8623**	.8812	.9088	.9238
1,000	.5856	.6186	.5089	**.4518**	.4723	.5007	.5134	.5150	.5213	.5449	.5132	.4826	.4563	.4591	.4893	.5277	.5534
2,000	.2762	.4266	.2939	**.2172**	.2262	.2392	.2473	.2455	.2485	.2588	.2449	.2317	.2223	.2288	.2560	.2951	.3319
5,000	.1122	.3285	.1850	.1002	.1044	.1078	.1115	.1064	.1066	.1073	.1032	.1001	.0999	.1045	.1193	.1409	.1627
10,000	.0570	.2949	.1473	.0584	.0604	.0616	.0645	.0587	.0582	.0558	.0544	.0534	.0538	.0561	.0637	.0753	.0874

Note. Bold font type indicates the smallest value with $C \geq 4$ or $S \leq .5$ at a given sample size less than 2,000; Number in shaded cell means that smoothing increases total equating error at a certain sample size and a certain smoothing parameter.

Table 13

Average Mean Squared Errors of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	1.5190	1.0167	1.1252	1.2195	1.2630	1.3270	1.3649	1.3972	1.4145	1.3951	1.2876	1.1917	1.0807	1.0296	**.9952**	.9855	.9822
200	.7499	.5049	.5643	.5966	.6206	.6467	.6645	.6810	.6903	.6927	.6411	.5978	.5499	.5311	**.5250**	.5288	.5327
500	.2955	.2179	.2414	.2416	.2489	.2547	.2631	.2693	.2711	.2769	.2590	.2453	.2322	**.2291**	.2336	.2447	.2539
1,000	.1457	.1270	.1384	.1260	.1284	.1276	.1308	.1333	.1344	.1379	.1301	.1251	**.1220**	.1234	.1315	.1435	.1542
2,000	.0760	.0865	.0920	.0737	.0735	.0695	.0712	.0718	.0717	.0743	.0710	.0693	**.0691**	.0705	.0755	.0834	.0919
5,000	.0277	.0548	.0565	.0356	.0336	.0273	.0282	.0276	.0270	.0292	.0281	.0278	.0283	.0292	.0315	.0348	.0386
10,000	.0151	.0479	.0487	.0267	.0242	.0172	.0176	.0168	.0159	.0176	.0172	.0171	.0175	.0181	.0195	.0212	.0231

Note. Bold font type indicates the smallest value with $C \geq 4$ or $S \leq .5$ at a given sample size less than 2,000; Number in shaded cell means that smoothing increases total equating error at a certain sample size and a certain smoothing parameter.

Table 14

Average Mean Squared Errors of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing (*C*)								Cubic Spline Postsmoothing (*S*)							
		2	3	4	5	6	7	8	9	.01	.05	.1	.2	.3	.5	.75	1.0
100	1.5916	1.1242	1.1653	1.2598	1.3138	1.3953	1.4285	1.4700	1.4916	1.4540	1.3353	1.2332	1.1167	1.0702	**1.0478**	1.0405	1.0387
200	.7773	.5829	.5618	.6063	.6391	.6739	.6885	.7081	.7160	.7117	.6572	.6056	.5480	**.5302**	.5319	.5396	.5446
500	.2976	.2788	.2209	.2330	.2445	.2564	.2625	.2689	.2727	.2754	.2565	.2386	.2211	**.2196**	.2313	.2468	.2569
1,000	.1555	.1917	.1218	.1233	.1302	.1352	.1379	.1408	.1425	.1453	.1361	.1282	**.1215**	.1238	.1379	.1570	.1720
2,000	.0762	.1432	.0666	.0627	.0663	.0681	.0687	.0700	.0705	.0719	.0678	.0645	**.0617**	.0629	.0710	.0840	.0971
5,000	.0322	.1155	.0356	.0288	.0306	.0303	.0301	.0307	.0306	.0311	.0298	.0290	.0285	.0292	.0327	.0385	.0448
10,000	.0163	.1055	.0242	.0168	.0180	.0170	.0164	.0166	.0164	.0162	.0159	.0157	.0157	.0161	.0175	.0202	.0232

Note. Bold font type indicates the smallest value with $C \geq 4$ or $S \leq .5$ at a given sample size less than 2,000; Number in shaded cell means that smoothing increases total equating error at a certain sample size and a certain smoothing parameter.

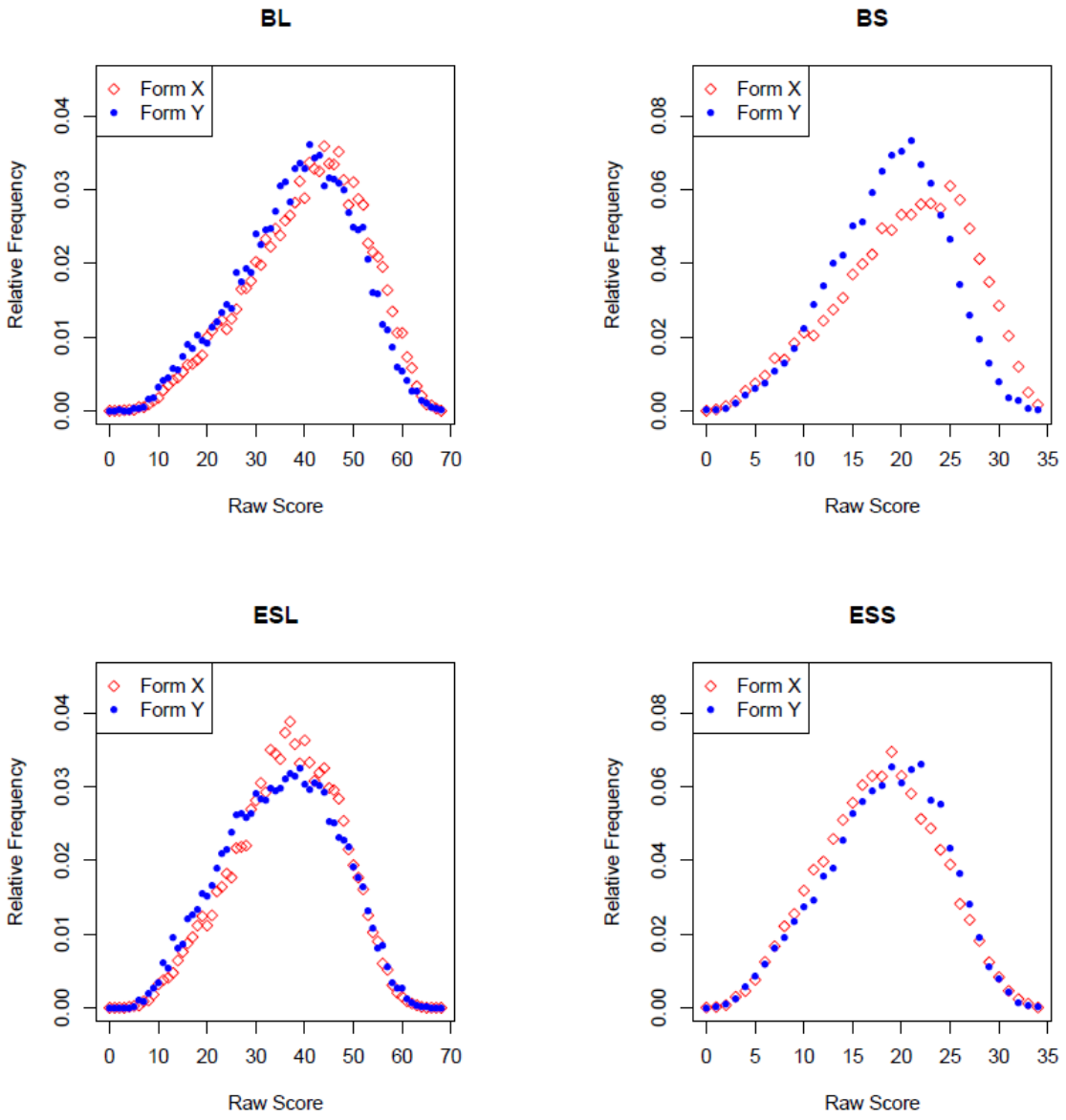

Figure 1. Distributions of pseudo-test forms for the whole dataset.

Chapter 9: Automated Selection of Smoothing Parameters in Equipercentile Equating

Chunyan Liu and Michael J. Kolen
The University of Iowa, Iowa City, IA

Abstract

Smoothing techniques are designed to improve the accuracy of equating functions. The main purpose of this study was to compare seven model selection strategies in selecting smoothing parameters for polynomial loglinear presmoothing (*C*) and cubic spline postsmoothing (*S*) for mixed-format tests for equiperecentile equating under a random groups design. For polynomial loglinear presmoothing, the *C* parameter was selected using the following model selection strategies: likelihood ratio chi-square test (G^2), Pearson chi-square test (*PC*), likelihood ratio chi-square difference test (G^2diff), Pearson chi-square difference test (*PCdiff*), Akaike Information Criterion (*AIC*), Bayesian Information Criterion (*BIC*), and Consistent Akaike Information Criterion (*CAIC*). For cubic spline postsmoothing, the *S* parameter was selected using the ± 1 standard error of equating (± 1 *SEE*) method. Four pseudo-test data were used to evaluate the performance of each model selection method and sample sizes of 100, 200, 500, 1,000, 2,000, 5,000, and 10,000 were investigated. The results indicate that among all model selection methods, either the *AIC* statistic in loglinear presmoothing or the ± 1 *SEE* method in cubic spline postsmoothing introduced the least amount of the average squared equating bias. For loglinear presmoothing, only the *AIC* and G^2diff statistics performed well in reducing the random errors across all investigated situations, and only the *AIC* statistic performed well in reducing the total equating error for $N \leq 5{,}000$ consistently, which suggests that the effectiveness of loglinear presmoothing in reducing the equating error depends on the model selection strategies. None of the smoothing techniques always worked well for sample size of 10,000. In addition, the results indicated that, considering all sample sizes and equating conditions, postsmoothing with the ± 1 *SEE* method performed better than any model selection strategy in loglinear presmoothing in reducing the random error and total error of equipercentile equating.

Automated Selection of Smoothing Parameters in Equipercentile Equating

Alternate forms are often constructed in testing organizations for test security purposes. No matter how much time and effort is spent by test developers to construct test forms that are as similar as possible, the resulting alternate forms are somewhat different in difficulty. Therefore, equating is a very important step in the test scoring process. According to Kolen and Brennan (2004), 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."

When conducting equating, two sources of error are present: systematic error and random error. Systematic error (or equating bias) can be introduced in several ways, such as smoothing, violation of statistical assumptions, improperly implemented data collection design, etc. Random error, often indexed by the standard error of equating (*SEE*), arises whenever examinees are considered a sample from the population.

Smoothing, including presmoothing and postsmoothing, is designed to reduce random error without introducing too much systematic error. The observed score distributions are smoothed prior to conducting equating in presmoothing and the raw score equating results are smoothed directly in postsmoothing. The most commonly used smoothing methods are polynomial loglinear presmoothing (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Darroch & Ratcliff, 1972; Haberman, 1974a, 1974b, 1978; Hanson, 1990; Hanson, Zeng, & Colton, 1994; Holland & Thayer, 1987, 2000; Moses, 2009; Moses & Holland, 2009a, 2009b, 2009c, 2010; Rosenbaum & Thayer, 1987) and cubic spline postsmoothing (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Kolen, 1984; Kolen & Jarjoura, 1987; Hanson et al., 1994; Zeng, 1995). Using the fixed smoothing parameter procedures (same amount of smoothing was used from sample to sample), previous research (Colton, 1995; Cui & Kolen, 2008; Hanson et al., 1994) indicated that both polynomial loglinear presmoothing and cubic spline postsmoothing could improve equating accuracy and the results provided by presmoothing and postsmoothing were comparable. However, in Chapter 8, it was found that postsmoothing often led to slightly less equating error than presmoothing.

There are some limitations with the previous studies. First, the smoothing parameter is fixed. However, in practice, it is often the case that different degrees of smoothing are used for different observed score distributions and different equating conditions. Second, the optimal

smoothing parameter is never known because the optimal smoothing parameter is dependent on several factors, such as sample size, number of score points, and population score distributions.

The primary purpose of this study is to compare the performance of loglinear presmoothing and cubic spline postsmoothing with the smoothing parameters selected by different model selection strategies, which allow the smoothing parameters to vary from one equating to another. For the loglinear presmoothing, the smoothing parameter was selected using seven model selection statistics: likelihood ratio chi-square (G^2) (Haberman, 1974a), likelihood ratio difference chi-square (G^2diff) (Haberman, 1974a), Pearson chi-square (*PC*), Pearson difference chi-square (*PCdiff*), Akaike Information Criterion (*AIC*) (Akaike, 1981), Bayesian Information Criterion (*BIC*) (Schwarz, 1978), and Consistent Akaike Information Criterion (*CAIC*) (Bozdogan, 1987). For cubic spline postsmoothing, the smoothing parameter using the method discussed in Kolen and Brennan (2004, Chapter 3), which is referred to the ± 1 standard error of equating (*SEE*) method in this chapter.

Smoothing Methods

Polynomial Loglinear Presmoothing

In polynomial loglinear presmoothing, the logarithm of the sample score frequency is fit using a polynomial function (Darroch & Ratcliff, 1972; Haberman, 1974a, 1974b, 1978; Holland & Thayer, 1987, 2000; Rosenbaum & Thayer, 1987), that is, for a particular score point *x*,

$$\log[N\hat{f}(x)] = \omega_0 + \omega_1 x + \omega_2 x^2 + ... + \omega_C x^C , \quad (1)$$

where N is the sample size, $\hat{f}(x)$ is the model implied relative frequency distribution, *C* is the degree of polynomial typically chosen from 2 to 10, and ω_0, ω_1, …, and ω_C are the estimated parameters for the polynomial function. The Newton-Raphson algorithm is typically used to compute the maximum likelihood estimates (*MLE*) of ω_0, ω_1, …, and ω_C (Haberman, 1974a; Holland & Thayer, 1987).

One desirable feature of the *MLE* of the parameters in Equation 1 is that the first *C* moments of the resulting fitted distribution are the same as those of the observed score distribution (Darroch & Ratcliff, 1972). For example, if *C*=4, the mean, standard deviation, skewness, and kurtosis are identical for the observed score distribution and the fitted score distribution.

One important consideration in using polynomial loglinear presmoothing is the selection of the *C* parameter. The simplest model that adequately fits the observed score distribution is preferred because more complicated models might introduce more equating error. The visual inspection of whether the fitted distribution adequately fits the observed score distribution well is always necessary. However, visual inspection provides little help when comparing values of *C* that are close to one another (e.g., *C*=5 versus *C*=6). Other than visual inspection of the fitted distribution, several statistical methods can be used to assess the model fit.

Haberman (1974a) stated that if L_i^2 is the likelihood ratio chi-square, or G^2, for the maximum likelihood fit of the model with a degree of freedom (*df*) *i*, then $L_{i-1}^2 - L_i^2$ is a likelihood ratio difference chi-square with a *df* of 1. A procedure for model selection using this difference chi-square was also reviewed by Hanson (1990) and Kolen and Brennan (2004, pp. 74-75). Holland and Thayer (1998, 2000) also suggested the Pearson chi-square (*PC*) statistic for univariate loglinear presmoothing. These two statistics are defined in Equation 2 and 3:

$$G^2 = 2\sum_j Nf(x_j)\log_e\left(\frac{f(x_j)}{\hat{f}(x_j)}\right), \tag{2}$$

$$PC = N\sum_j \frac{(f(x_j) - \hat{f}(x_j))^2}{\hat{f}(x_j)}, \tag{3}$$

where $f(x_j)$ is the observed score distribution, and the summation is over all score points. The degrees of freedom for these chi-square statistics are *K*-*C*, where *K* is the number of distinct score points minus 1.

These chi-square statistics are typically used in two ways in selecting the smoothing parameter in polynomial loglinear presmoothing. The first is to use the difference between the overall chi-square statistics for the model with parameter *C*+1 and the model with parameter *C* (Haberman, 1974a; Hanson, 1990; Kolen & Brennan, 2004; Moses & Holland, 2009c). The difference chi-square statistics test whether the model with parameter *C*+1 fits the data better than the model with parameter *C*. A significant difference indicates a better fit using the higher degree polynomial. The second is to use the overall goodness-of-fit statistic to test whether the model with parameter *C* fits the observed data adequately (Cui, 2006; Kolen & Brennan, 2004; Moses & Holland, 2009c). A non-significant chi-square statistic suggests a good fit. However,

Kolen and Brennan (2004, p. 80) indicated that these significance tests only provided a guide for model selection rather than a definitive decision rule.

Moses and Holland (2009c) introduced the following statistics that took parsimony into account in selecting the smoothing parameter for univariate loglinear smoothing models: *AIC*, *BIC*, and *CAIC*. These parsimony statistics are defined to balance the overall fit of the model (G^2) and the parameterization (i.e., *C*) (Moses & Holland, 2009c). The model selected using any of these statistics is the model that has the smallest value of the parsimony statistic. These statistics are defined in Equations 4 through 6:

$$AIC = G^2 + 2(C+1), \tag{4}$$

$$BIC = G^2 + \log_e(N)(C+1), \tag{5}$$

$$CAIC = G^2 + [1+\log_e(N)](C+1). \tag{6}$$

In this study, the following statistics were investigated for selecting the smoothing parameters for polynomial loglinear presmoothing: G^2 overall chi-statistic (G^2), *PC* overall chi-square statistic (*PC*), G^2 difference chi-statistic (G^2diff), *PC* difference chi-square statistic (*PCdiff*), *AIC*, *BIC*, and *CAIC*.

Cubic Spline Postsmoothing

Cubic spline smoothing was first described by Reinsch (1967), and the spline fitting algorithm was presented by de Boor (1978, pp. 235-243). This procedure was first applied in the field of equating by Kolen (1984) and is summarized as follows.

Given integer scores $x_0, x_1, ..., x_k$ and $x_0 < x_1 < ... < x_k$ where *k* is the maximum score point, the cubic spline function defined for the interval $x_i < x < x_{i+1}$ is

$$\hat{d}_Y(x) = v_{0i} + v_{1i}(x - x_i) + v_{2i}(x - x_i)^2 + v_{3i}(x - x_i)^3, \tag{7}$$

where the coefficients v_{0i}, v_{1i}, v_{2i}, and v_{3i} are constants and allowed to change from one interval to the next. At each score point, x_i, the second derivative of the cubic spline function is continuous. Different cubic spline functions are estimated for each interval of $[x_i, x_i + 1]$ over the range of score points x_{low} to x_{high}, where x_{low} is the lowest score in the range, x_{high} is the highest score point in the range, and $0 \le x_{low} \le x \le x_{high} \le K$.

The cubic spline function is then obtained by requiring minimum curvature and satisfying the following inequality:

$$\frac{\sum_{i=low}^{high}\left[\frac{\hat{d}_Y(x_i)-\hat{e}_Y(x_i)}{\widehat{se}\left[\hat{e}_Y(x_i)\right]}\right]^2}{x_{high}-x_{low}+1}\le S, \tag{8}$$

where the summation is over the range of scores that is used to fit the spline function. In the context of equipercentile equating, $\hat{e}_Y(x_i)$ is the equated raw score of x_i on Form X; $\hat{d}_Y(x_i)$ is the equated score of x_i after smoothing; and $\widehat{se}\left[\hat{e}_Y(x_i)\right]$ is the estimated standard error of the equated raw score.

Parameter S (≥ 0) is the smoothing parameter for cubic spline postsmoothing, which controls the degree of smoothing. S=0 means no smoothing, where the smoothed equivalents equal the unsmoothed equivalents. If S is very large, the spline function is a straight line. Several studies (Colton, 1995; Cui, 2006; Cui & Kolen, 2009; Hanson et al., 1994; Kolen, 1991; Kolen & Brennan, 2004; Kolen & Jarjoura, 1987; Zeng, 1995) suggested that a value of S between 0 and 1 yields appropriate results in equating.

One concern with cubic spline postsmoothing is that no appropriate statistical tests are available for deciding which S value should be chosen in practice. The S value is typically selected by examining the equating relationship visually. Kolen and Brennan (2004) recommended that the maximum S value be chosen such that the smoothed equated scores are within ± 1 standard errors of equating across the score points with percentile ranks between .5 and 99.5, and a linear interpolation procedure is used to obtain the smoothed equated score outside of this range, which is referred to the ± 1 *SEE* method in this study.

Method

Detailed information about the tests, construction of the pseudo-test forms, population score distributions, population equating relationships, and the evaluation criteria are not provided in this paper and can be found in Chapter 8. The data used for this study are from the College Board Advanced Placement Program® (AP®) tests: Biology 2005 (BL), Biology 2006 (BS), Environmental Science 2005 (ESL), and Environmental Science 2006 (ESS). The distributions of the pseudo-test forms based on the whole dataset are considered as the population score distributions from which the true equating relationship are derived. The evaluation criteria are

average squared bias (*ASB*), average squared error (*ASE*), average mean squared error (*AMSE*), and conditional equating bias (*Bias*), standard errors of equating (*SEE*), and root mean square errors of equating (*RMSE*) at each score point. However, the sampling procedures are slightly different and provided as follows:

Sampling Procedures

The following steps were applied to conduct the data analyses for all four pseudo-tests:

1. Random samples of size *N* were drawn from the population distributions separately for Form X and Form Y.
2. Equipercentile equating was conducted using the two sample distributions.
3. Cubic spline postsmoothing was applied to the equating results obtained in Step 2 with the *S* parameter selected using the ± 1 *SEE* method.
4. The distributions for the two samples obtained in Step 1 were smoothed using polynomial loglinear presmoothing with the *C* parameter selected using G^2, G^2*diff*, *PC*, *PCdiff*, *AIC*, *BIC*, and *CAIC*.
5. Equipercentile equating was conducted using the smoothed distributions obtained in Step 4.
6. Steps 1 to 5 were repeated 500 times for each sample size (*N*=100, 200, 500, 1,000, 2,000, 5,000, and 10,000) and each test (BL, BS, ESL, and ESS).

All analyses were conducted using the open source *Equating Recipes* (*ER*; Brennan, Wang, Kim, & Seol, 2009) C code.

Results

In this section, the comparison of average squared bias yielded by different model selection methods are presented first, which is followed by the comparisons of average squared error, and average mean squared error. Next, the equating bias, standard errors of equating, and root mean square errors at each score point were compared among G^2, G^2*diff*, *AIC* in polynomial loglinear presmoothing, and the ±1 *SEE* method in cubic spline postsmoothing for the ESS pseudo-test at sample sizes of 200, 500, 1,000, and 2,000. Only the G^2, G^2*diff*, *AIC*, and the ±1 *SEE* methods were provided because 1) they are representative of the other model selection statistics, and 2) it was easier to observe the differences among different model selection procedures. The conditional results based on other pseudo-tests are similar.

Average Squared Bias

The *ASB* statistic indexes systematic error, or bias, in equating. *ASB* for the equating results without smoothing (denoted as "none"), with loglinear presmoothing (the *C* parameter was selected using G^2, G^2*diff*, *PC*, *PCdiff*, *AIC*, *BIC*, and *CAIC*), and with cubic spline postsmoothing (the *S* parameter was selected using the ± 1 *SEE* method) for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are provided in Tables 1 through 4, respectively. For the results without smoothing, equating bias tended to decrease with the increase of the sample sizes across all pseudo-tests.

For polynomial loglinear presmoothing, *ASB* tended to decrease as sample size increased for most situations. Compared to no smoothing, all model selection methods in loglinear presmoothing had larger *ASB* for sample sizes larger than 500. For a sample size of 100, loglinear presmoothing tended to reduce *ASB* compared to no smoothing. For example, for the BS test at sample size of 100, *ASB* for the G^2*diff* and *AIC* methods were .0085 and .0057 respectively, which were smaller than .0094 that was the *ASB* without smoothing. In addition, the *AIC* method had the smallest *ASB* among all model selection methods across all sample sizes and all pseudo tests except for the BL test at sample size of 100.

For cubic spline postsmoothing, *ASB* decreased as the sample sizes increased across all sample sizes and all pseudo-tests. Similar to the results found for loglinear presmoothing, cubic spline postsmoothing did not necessarily have larger *ASB* than no smoothing, especially for smaller sample sizes ($N \leq 200$). For example, for the BL test, *ASB* was smaller for cubic spline postsmoothing than for no smoothing for $N \leq 200$.

In general, either the *AIC* method in loglinear presmoothing or the ± 1 *SEE* method in cubic spline postsmoothing introduced the least amount of *ASB* among all model selection methods as indicated by the bold *ASB* values in Tables 1 through 4.

Average Squared Error

The *ASE* statistic indexes random error in equating. *ASE* of equating without smoothing (denoted as "none"), with loglinear presmoothing (the *C* parameter was selected using G^2, G^2*diff*, *PC*, *PCdiff*, *AIC*, *BIC*, and *CAIC*), and with cubic spline postsmoothing (the *S* parameter was selected using the ± 1 *SEE* method) for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are provided in Tables 5 through 8, respectively. The smallest *ASE* in each row is in bold.

Similar to the results for *ASB*, *ASE* without smoothing tended to decrease as sample size increased for all pseudo-test forms.

For polynomial loglinear presmoothing, *ASE* tended to decrease significantly as sample size increased for all model selection methods across all investigated equating conditions. However, only G^2*diff* and *AIC* model selection methods consistently reduced *ASE* across all equating conditions. The performance of G^2, *PC*, *PCdiff*, *BIC*, and *CAIC* was dependent on sample sizes and pseudo-tests. For example, for the BL test, the *ASEs* for the G^2 model selection method are 4.2245, 2.3751, 1.3947, .7534, .2793, .1279, and .0541 for sample sizes of 100, 200, 500, 1,000, 2,000, 5,000, and 10,000, respectively. The corresponding *ASE*s without smoothing are 6.5650, 3.2520, 1.2015, .6210, .3149, .1344, and .0613. So the G^2 model selection method reduced *ASE* for sample sizes of 100, 200, 2,000, 5,000, and 10,000 but not for the sample sizes of 500 and 1,000. Similar results were found for *PC*, *PCdiff*, *BIC*, and *CAIC* at different equating conditions.

For the ± 1 *SEE* method in cubic spline postsmoothing, *ASE* tended to decrease substantially as sample size increased across all investigated equating conditions, which is similar to the results of loglinear presmoothing. However, the cubic spline postsmoothing reduced the random equating error consistently across all equating conditions.

In general, the ± 1 *SEE* method in cubic spline postsmoothing tended to produce smaller *ASE* than any model selection methods in loglinear presmoothing especially when the sample size was not too small ($N \geq 500$). When sample sizes are less than 200, the G^2 and *PC* model selection methods for loglinear presmoothing had smaller values of *ASE* than ± 1 *SEE* method in cubic spline postsmoothing. The ± 1 *SEE* method in cubic spline postsmoothing almost always yielded smaller *ASE* (except for the ESL test at sample size of 10,000) compared to the loglinear presmoothing with the *C* parameter selected by G^2*diff* and *AIC* model selection methods which are the methods that worked well consistently in reducing the average squared equating error across all equating conditions.

Average Mean Squared Error

The *AMSE* statistic indexes overall equating error, including both random and systematic components. *AMSE* for equating without smoothing (denoted as “none”), with loglinear presmoothing (the *C* parameter was selected using G^2, G^2*diff*, *PC*, *PCdiff*, *AIC*, *BIC*, and *CAIC*), and with cubic spline postsmoothing (the *S* parameter was selected using the ± 1 *SEE* method)

for BL, ESL, BS, and ESS pseudo-tests at different sample sizes are provided in Tables 9 through 12, respectively. Again, the smallest value in each row is in bold. For all results with and without smoothing, *AMSE* tended to decrease as sample size increased, across all pseudo-tests because the total equating error is dominated by the random equating error.

For polynomial loglinear presmoothing, none of the model selection methods reduced *AMSE*, compared to no smoothing, across all sample sizes and all pseudo-tests for a sample size of 10,000. For example, for the BL test, only the G^2, *PC*, and *AIC* model selection methods reduced *AMSE* compared to no smoothing, while for the ESL, BS, and ESS tests, none of the model selection methods reduced *AMSE* compared to no smoothing. For a sample size of 5,000 or less, among all presmoothing methods, only the *AIC* model selection method reduced *AMSE* compared to no smoothing across all equating conditions and the performance of other model selection methods was dependent on sample size and pseudo-test. For example, compared to no smoothing, the G^2*diff* model selection method did not reduce *AMSE* for the sample size of 500 and 5,000 for BL and BS tests, consistently reduced *AMSE* for the ESL test, and did not reduce *AMSE* for the sample size of 5,000 for the ESS test. Similar results were found for the rest of the model selection methods.

The ± 1 *SEE* method for cubic spline postsmoothing reduced *AMSE*, compared to no smoothing, across all equating conditions except for the BS test at sample sizes of 5,000 and 10,000 where cubic spline postsmoothing increased *AMSE* compared the those without smoothing (.0285 and .0175 with postsmoothing versus .0277 and .0151 without smoothing for N=5,000 and 10,000, respectively).

In general, the ± 1 *SEE* method in cubic spline postsmoothing tended to produce smaller *AMSE* than any model selection methods in loglinear presmoothing especially when the sample size was not too small ($N \geq 500$). The G^2 and *PC* model selection methods in loglinear presmoothing tended to have the smallest *AMSE* for sample sizes of 200 or less. Across all sample sizes and all pseudo-tests, the ± 1 *SEE* method in cubic spline postsmoothing almost always yielded smaller average mean squared equating error (except for the BS test at sample sizes of 5,000 and 10,000) compared to the loglinear presmoothing with the *C* parameter selected by the *AIC* model selection method which is the method that worked well in reducing *AMSE* for $N \leq 5{,}000$ across all pseudo-tests.

Conditional Results Comparison between Presmoothing and Postsmoothing

Figure 1 shows the equating bias of ESS pseudo-test at each score point for sample sizes of 200, 500, 1,000, and 2,000. As shown in the figure, the biases were smaller for the *AIC* strategy and larger for the G^2 statistic at most score points. The ±1 *SEE* method in postsmoothing tended to produce larger equating bias than the *AIC* statistic. Also, smoothing tended to increase the bias at most score points. It seemed that the equating bias tended to decrease as the sample size increased, which was consistent with the findings from Table 4.

Figure 2 shows the standard error of equating of ESS pseudo-test at each score point for different sample sizes. Across all sample sizes, the standard errors of equating were the smallest for the ±1 *SEE* method in postsmoothing at most score points. The G^2 statistic tended to yield a smaller standard error of equating than *AIC* and G^2*diff* at most score points for *N*=200. The standard errors of equating yielded by *AIC* and the ±1 *SEE* method tended to be smaller than those without smoothing across most score points; however, the G^2 and G^2*diff* statistic yielded larger standard errors of equating than the unsmoothed equating results at many score points, especially for *N*=1,000, which was consistent with the results in Table 8.

Figure 3 shows the root mean square errors of equating of ESS test at each score point for different sample sizes. Across all sample sizes, the root mean square errors of equating were the smallest for the ±1 *SEE* method in postsmoothing at most score points. Both *AIC* in presmoothing and the ±1 *SEE* method in postsmoothing tended to yield smaller root mean square errors of equating than the unsmoothed equating results across almost all score points. However, the G^2 statistic yielded larger root mean square errors of equating than the unsmoothed equating at many score points, especially for *N*=500 and 1,000. For example, for *N*=1,000, the root mean square errors were larger for the G^2 statistic than the unsmoothed equating results at score points of 0 through 5, 10 through 15, 20 through 24, and 28 through 34, which supported the findings based on the *AMSE* of equipercentile equating in Table 12. For *N*=2,000, all procedures tended to reduce the root mean square errors at most of the score points except at very low or very high score ranges.

Discussion and Conclusions

The primary purpose of this study was to compare the polynomial loglinear presmoothing (the *C* parameter was selected from 2 through 9 using G^2, G^2*diff*, *PC*, *PCdiff*, *AIC*, *BIC*, and *CAIC*) and cubic spline postsmoothing (the *S* parameter was selected .01, .05, .1, .2, .3, .5, .75,

and 1.0 using the ± 1 *SEE* method) with a wide range of sample sizes (*N*=100, 200, 500, 1,000, 2,000, 5,000, and 10,000) under the random groups design. The main difference between this study and previous studies is that the smoothing parameters (*C* and *S*) were allowed to vary from one sample to another.

The results indicate that among all model selection methods, the *AIC* statistic in loglinear presmoothing and the ± 1 *SEE* method in cubic spline postsmoothing introduced the least amount of bias (as indexed by *ASB*) of equipercentile equating across all equating conditions. One general conclusion from previous research is that presmoothing can reduce the random errors of equating (as indexed by *ASE*) and total equating error (as indexed by *AMSE*), using a fixed-*C* procedure (Colton, 1995; Cui & Kolen, 2006, 2009; Hanson et al., 1994; Kolen & Brennan, 2004). However, it was found in this study that among the seven model selection strategies in loglinear presmoothing, only the *AIC* and G^2*diff* statistics performed well in reducing the random errors (as indexed by *ASE*) in all investigated situations, and only the *AIC* statistic performed well in consistently reducing the total equating error (as indexed by *AMSE*) for $N \leq 5{,}000$, which suggests that the effectiveness of loglinear presmoothing in reducing the equating error is dependent on the model selection strategy used. An ineffective model selection strategy might introduce equating error compared to no smoothing.

This result might be due to the following: 1) The *C* parameters selected by the same model selection strategy can be quite different for Form X and Form Y, and 2) the selected *C* values can be quite different from one sample to another. Table 13 provides a bivariate distribution of the *C* parameters for Form X and Form Y selected by the *AIC* statistic in loglinear presmoothing based on the 500 pairs of random samples for the BS test at the sample size of 500. Clearly, different *C* parameters were often selected for Form X and Form Y, and the selected *C* parameters ranges from 2 to 9 for both Form X and Form Y. The bivariate distributions of the *C* parameters selected by other statistics were very similar.

Another general conclusion about smoothing from previous research is that presmoothing and postsmoothing methods tended to provide comparable results in reducing equating error using fixed smoothing parameter procedures (Colton, 1995; Cui & Kolen, 2006, 2009; Hanson et al., 1994). However, the results in this study suggest that, considering all sample sizes and equating conditions, postsmoothing with the ± 1 *SEE* method tends to be better than any model selection strategy in loglinear presmoothing in reducing the random error and total error of

equipercentile equating, which is consistent with the results found in Chapter 8. This finding probably due to the reason that postsmoothing directly smooths the equating function whereas presmoothing smooths the observed score distributions.

Comparing with the results in Chapter 8, it was found that the fixed smoothing parameter procedures with medium amount of smoothing (such as, C=5 or 6 and S=.2 or .3) might perform better than the *AIC* statistic and the ± 1 *SEE* method in reducing *AMSE*, especially for sample size of 2000 or less. For example, *AMSE* of the ESS test are 1.314, .639, .244, .130, .066, and .031 for C=5, and 1.4239, .6941, .2664, .1406, .0713, and .0308 for the *AIC* statistic for sample size of 100, 200, 500, 1,000, 2,000, and 5,000, respectively. Similar results were found for cubic spline postsmoothing. However, this does not mean that the fixed smoothing parameter procedure should be preferred because in practice, it is very common that different smoothing parameters are used for different equating conditions.

When using smoothing in practice, it is recommended that the best methods considered in this study be used (the *AIC* method in loglinear presmoothing and the ± 1 *SEE* method in cubic spline postsmoothing) to help choose a smoothing parameter or parameters to consider further. The chosen parameter or parameters should be used as a starting place for equating. All of the other characteristics suggested by Kolen and Brennan (2004, Chapter 3) for choosing smoothing parameters should also be considered, including inspection of graphs of smoothed and unsmoothed distributions and equating relationships, comparing moments of equated raw and scale scores, and inspection of conversion tables when conducting operational equipercentile equating.

References

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

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, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Colton, D. A. (1995). *Comparing smoothing and equating methods using small sample sizes*. Unpublished Ph.D. Dissertation. University of Iowa, Iowa City, Iowa.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: the cubic B-spline presmoothing method and the direct presmoothing*. Unpublished doctoral dissertation, University of Iowa, Iowa City, Iowa.

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32,* 334-347

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

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics, 43*, 1470-1480.

De Boor, C. (1978). *A practical guide to splines* (Applied Mathematical Science, Volume 27). New York: Springer-Verlag.

Haberman, S. J. (1974a). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30*, 589-600.

Haberman, S. J. (1974b). *The analysis of frequency data*. Chicago: University of Chicago.

Haberman, S. J. (1978). *Analysis of qualitative data. Vol. 1. Introductory topics*. New York: Academic.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions*. (ACT Research Report 90-4). Iowa City, IA: American College Testing.

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

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Technical Report 87-89). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1998). *Univariate and bivariate loglinear models for discrete test score distributions*. (Technical Report 98-1). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133-183.

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

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28,* 257-272.

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

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equatingunder the common item nonequivalent populations design. *Psychometrika, 52,* 43-59.

Moses, T. (2009). A comparison of statistical significance tests for selecting equating functions. *Applied Psychological Measurement, 33,* 285-306.

Moses, T., & Holland, P. W. (2009a). *Selection strategies for bivariate loglinear smoothing models and their effects on NEST equating functions*. (Technical Report 09-04). Princeton, NJ: Educational Testing Service.

Moses, T., & Holland, P. W. (2009b). *Alternative loglinear smoothing models and their effect on equating function accuracy*. (Technical Report 09-48). Princeton, NJ: Educational Testing Service.

Moses, T., & Holland, P. W. (2009c). Selection strategies for Univariate loglinear smoothing models and their effects on equating function accuracy. *Journal of Educational Measurement. 46,* 159-176.

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

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10*, 177-183.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British journal of Mathematical and Statistical Psychology, 40*, 43-49.

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

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement, 19,* 177-190.

Table 1

Average Squared Bias of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	.0774	.0663	.0646	.0600	.0594	.0712	.0611	.0589	**.0502**
200	.0255	.0378	.0373	.0530	.0332	.0303	.0519	.0500	**.0236**
500	.0077	.0268	.0257	.0406	.0356	**.0130**	.0447	.0494	.0134
1,000	.0029	.0313	.0290	.0249	.0217	**.0087**	.0282	.0324	.0108
2,000	.0019	.0213	.0251	.0172	.0163	.0088	.0213	.0238	**.0087**
5,000	.0005	.0092	.0093	.0144	.0124	.0073	.0176	.0186	**.0066**
10,000	.0002	.0063	.0064	.0118	.0102	.0065	.0162	.0170	**.0051**

Note. Bold font type indicates the smallest value at a given sample size.

Table 2

Average Squared Bias of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	.0975	.2668	.2464	.1360	.0828	.0565	.1536	.1797	**.0496**
200	.0266	.2330	.2347	.0733	.0409	**.0143**	.0698	.1031	.0259
500	.0107	.1195	.1603	.0164	.0252	.0147	.0161	.0224	**.0123**
1,000	.0049	.0533	.0596	.0140	.0151	.0135	.0149	.0148	**.0105**
2,000	.0025	.0148	.0201	.0144	.0144	.0124	.0149	.0150	**.0082**
5,000	.0005	.0124	.0126	.0135	.0139	.0111	.0156	.0156	**.0052**
10,000	.0002	.0093	.0094	.0125	.0129	.0102	.0146	.0150	**.0037**

Note. Bold font type indicates the smallest value at a given sample size.

Table 3

Average Squared Bias of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	.0094	.0231	.0293	.0085	.0115	**.0057**	.0103	.0168	.0266
200	.0025	.0145	.0207	.0095	.0066	**.0060**	.0097	.0115	.0195
500	.0013	.0294	.0208	.0251	.0295	**.0058**	.0312	.0361	.0129
1,000	.0008	.0210	.0307	.0085	.0164	**.0041**	.0136	.0156	.0096
2,000	.0002	.0079	.0121	.0050	.0056	**.0028**	.0109	.0120	.0066
5,000	.0001	.0029	.0033	.0034	.0037	**.0025**	.0068	.0077	.0048
10,000	.0000	.0022	.0022	.0027	.0029	**.0022**	.0044	.0050	.0039

Note. Bold font type indicates the smallest value at a given sample size.

Table 4

Average Squared Bias of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	.0138	.0764	.0781	.0494	.0348	**.0131**	.0377	.0535	.0198
200	.0038	.0554	.0625	.0265	.0256	**.0043**	.0209	.0347	.0109
500	.0011	.0102	.0209	.0066	.0072	**.0032**	.0179	.0183	.0059
1,000	.0005	.0064	.0068	.0036	.0063	**.0026**	.0153	.0212	.0049
2,000	.0002	.0034	.0048	.0032	.0033	**.0023**	.0040	.0045	.0027
5,000	.0002	.0024	.0025	.0030	.0030	**.0020**	.0035	.0035	.0022
10,000	.0000	.0017	.0017	.0026	.0026	.0017	.0031	.0032	**.0017**

Note. Bold font type indicates the smallest value at a given sample size.

Table 5

Average Squared Errors of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	6.5650	**4.2245**	4.3862	5.2892	5.9536	5.9651	5.3497	5.0838	4.8310
200	3.2520	2.3751	**2.3398**	3.1694	3.2126	2.9171	3.1467	3.0813	2.4001
500	1.2015	1.3947	1.2995	1.1737	1.3227	1.0417	1.1818	1.2876	**.9103**
1,000	.6210	.7534	.8977	.5282	.5499	.5484	.4988	.4979	**.4860**
2,000	.3149	.2793	.3052	.2870	.2852	.2782	.2717	.2667	**.2529**
5,000	.1344	.1279	.1317	.1219	.1285	.1195	.1291	.1309	**.1146**
10,000	.0613	.0541	.0545	.0538	.0548	.0542	.0538	.0545	**.0528**

Note. Bold font type indicates the smallest value at a given sample size.

Table 6

Average Squared Errors of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	6.1151	**3.8686**	3.9352	4.8789	5.3904	5.4246	4.7758	4.4192	4.4595
200	2.8872	1.8605	**1.8577**	2.5753	2.6550	2.5333	2.5578	2.4558	2.1499
500	1.1602	1.0741	.9582	1.0518	1.1469	.9952	1.0995	1.2084	**.9146**
1,000	.5807	.6824	.7878	.4846	.5181	.5056	.4629	.4643	**.4614**
2,000	.2736	.2270	.2831	.2408	.2378	.2392	.2212	**.2149**	.2227
5,000	.1117	.1054	.1076	.0980	.1042	.0978	.1067	.1062	**.0953**
10,000	.0568	.0501	.0508	.0497	.0505	**.0497**	.0515	.0534	.0500

Note. Bold font type indicates the smallest value at a given sample size.

Table 7

Average Squared Errors of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	1.4805	1.0322	**1.0132**	1.1792	1.2490	1.3805	1.1924	1.1088	1.0346
200	.7411	.5837	.5672	.6744	.6890	.6819	.6736	.6444	**.5349**
500	.2942	.3225	.3237	.2937	.3080	.2627	.2906	.3063	**.2256**
1,000	.1449	.1547	.1685	.1329	.1462	.1290	.1320	.1368	**.1152**
2,000	.0758	.0720	.0803	.0691	.0726	.0697	.0646	.0639	**.0636**
5,000	.0276	.0260	.0263	.0252	.0253	.0250	.0253	.0251	**.0237**
10,000	.0150	.0140	.0141	.0139	.0142	.0139	.0146	.0144	**.0136**

Note. Bold font type indicates the smallest value at a given sample size.

Table 8

Average Squared Errors of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	1.5902	1.0960	**1.0890**	1.2520	1.3118	1.4108	1.2154	1.1636	1.1347
200	.7735	.5643	**.5463**	.6588	.6522	.6898	.6343	.6026	.5502
500	.2965	.2875	.2674	.2872	.3013	.2633	.2752	.2818	**.2244**
1,000	.1549	.1607	.1756	.1340	.1463	.1380	.1391	.1399	**.1207**
2,000	.0760	.0663	.0747	.0662	.0668	.0690	.0644	.0666	**.0605**
5,000	.0320	.0309	.0311	.0302	.0307	.0289	.0302	.0295	**.0266**
10,000	.0163	.0152	.0156	.0149	.0157	.0148	.0166	.0171	**.0140**

Note. Bold font type indicates the smallest value at a given sample size.

Table 9

Average Mean Squared Errors of Equating for BL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	6.6425	**4.2908**	4.4508	5.3492	6.0130	6.0363	5.4108	5.1427	4.8812
200	3.2775	2.4130	**2.3770**	3.2224	3.2458	2.9474	3.1986	3.1313	2.4237
500	1.2092	1.4215	1.3252	1.2143	1.3583	1.0547	1.2264	1.3370	**.9237**
1,000	.6239	.7847	.9267	.5531	.5716	.5571	.5270	.5303	**.4968**
2,000	.3168	.3006	.3303	.3041	.3015	.2870	.2930	.2905	**.2616**
5,000	.1349	.1372	.1410	.1363	.1410	.1268	.1467	.1495	**.1212**
10,000	.0615	.0605	.0609	.0656	.0650	.0607	.0700	.0715	**.0579**

Note. Bold font type indicates the smallest value at a given sample size.

Table 10

Average Mean Squared Errors of Equating for ESL Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	6.2125	**4.1354**	4.1816	5.0150	5.4732	5.4811	4.9294	4.5988	4.5090
200	2.9138	2.0935	**2.0924**	2.6486	2.6959	2.5476	2.6276	2.5589	2.1758
500	1.1709	1.1936	1.1185	1.0682	1.1722	1.0098	1.1156	1.2308	**.9269**
1,000	.5856	.7357	.8474	.4986	.5331	.5191	.4778	.4791	**.4720**
2,000	.2762	.2418	.3032	.2552	.2522	.2516	.2361	**.2298**	.2309
5,000	.1122	.1178	.1201	.1115	.1181	.1088	.1223	.1218	**.1006**
10,000	.0570	.0594	.0602	.0622	.0634	.0599	.0661	.0684	**.0537**

Note. Bold font type indicates the smallest value at a given sample size.

Table 11

Average Mean Squared Errors of Equating for BS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	1.4900	1.0552	**1.0425**	1.1878	1.2605	1.3862	1.2026	1.1256	1.0612
200	.7436	.5982	.5879	.6839	.6956	.6879	.6834	.6559	**.5544**
500	.2955	.3519	.3445	.3188	.3375	.2685	.3218	.3424	**.2385**
1,000	.1457	.1758	.1992	.1414	.1627	.1331	.1455	.1524	**.1248**
2,000	.0760	.0800	.0924	.0742	.0782	.0726	.0755	.0760	**.0702**
5,000	.0277	.0289	.0296	.0286	.0290	**.0275**	.0321	.0328	.0285
10,000	.0151	.0161	.0162	.0166	.0171	.0162	.0190	.0194	.0175

Note. Bold font type indicates the smallest value at a given sample size.

Table 12

Average Mean Squared Errors of Equating for ESS Test

N	none	Polynomial Loglinear Presmoothing							Postsmoothing
		G^2	PC	G^2diff	PCdiff	AIC	BIC	CAIC	± 1 SEE
100	1.6039	1.1724	1.1671	1.3014	1.3465	1.4239	1.2531	1.2171	**1.1545**
200	.7773	.6197	.6088	.6854	.6778	.6941	.6552	.6373	**.5611**
500	.2976	.2977	.2883	.2939	.3085	.2664	.2931	.3000	**.2303**
1,000	.1555	.1670	.1824	.1376	.1526	.1406	.1544	.1611	**.1256**
2,000	.0762	.0696	.0795	.0695	.0701	.0713	.0684	.0711	**.0632**
5,000	.0322	.0332	.0336	.0333	.0337	.0308	.0337	.0330	**.0288**
10,000	.0163	.0169	.0173	.0175	.0182	.0165	.0198	.0203	**.0157**

Note. Bold font type indicates the smallest value at a given sample size.

Table 13

The C Parameters for Form X versus the C Parameters for Form Y Selected by the AIC Statistic Based on the 500 Pairs of Random Samples for BS Test at N=500

		Form Y								
		2	3	4	5	6	7	8	9	Total
Form X	2	0	0	0	0	1	0	0	0	1
	3	0	1	0	0	0	0	0	0	1
	4	10	17	157	20	27	3	24	6	264
	5	1	2	9	0	1	0	3	1	17
	6	4	8	83	15	12	4	17	5	148
	7	0	3	17	2	5	2	6	1	36
	8	2	0	11	3	2	0	4	2	24
	9	0	0	5	0	1	0	3	0	9
	Total	17	31	282	40	49	9	57	15	500

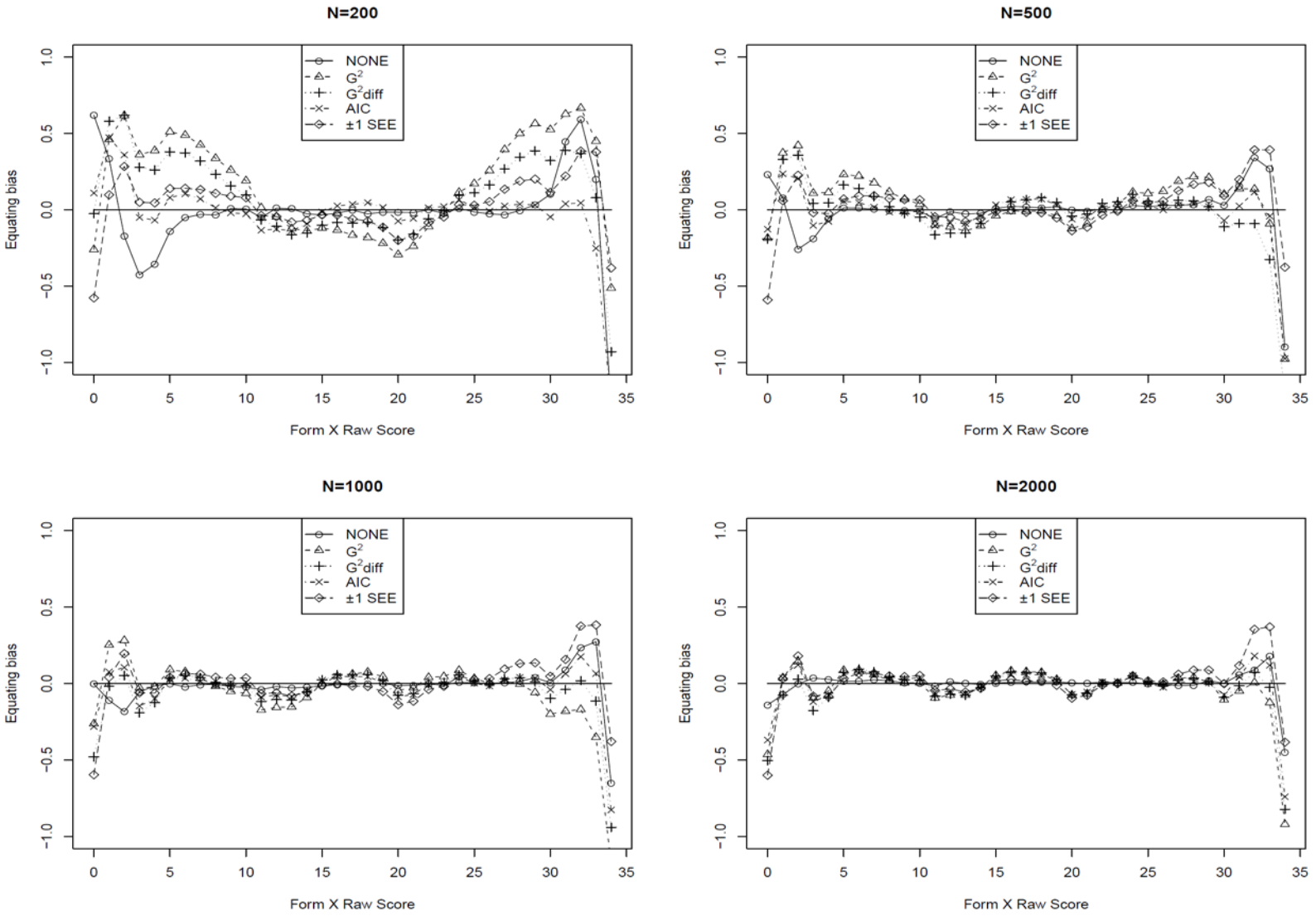

Figure 1. Equating bias for ESS pseudo-test with different smoothing strategies at *N*=200, 500, 1,000, and 2,000.

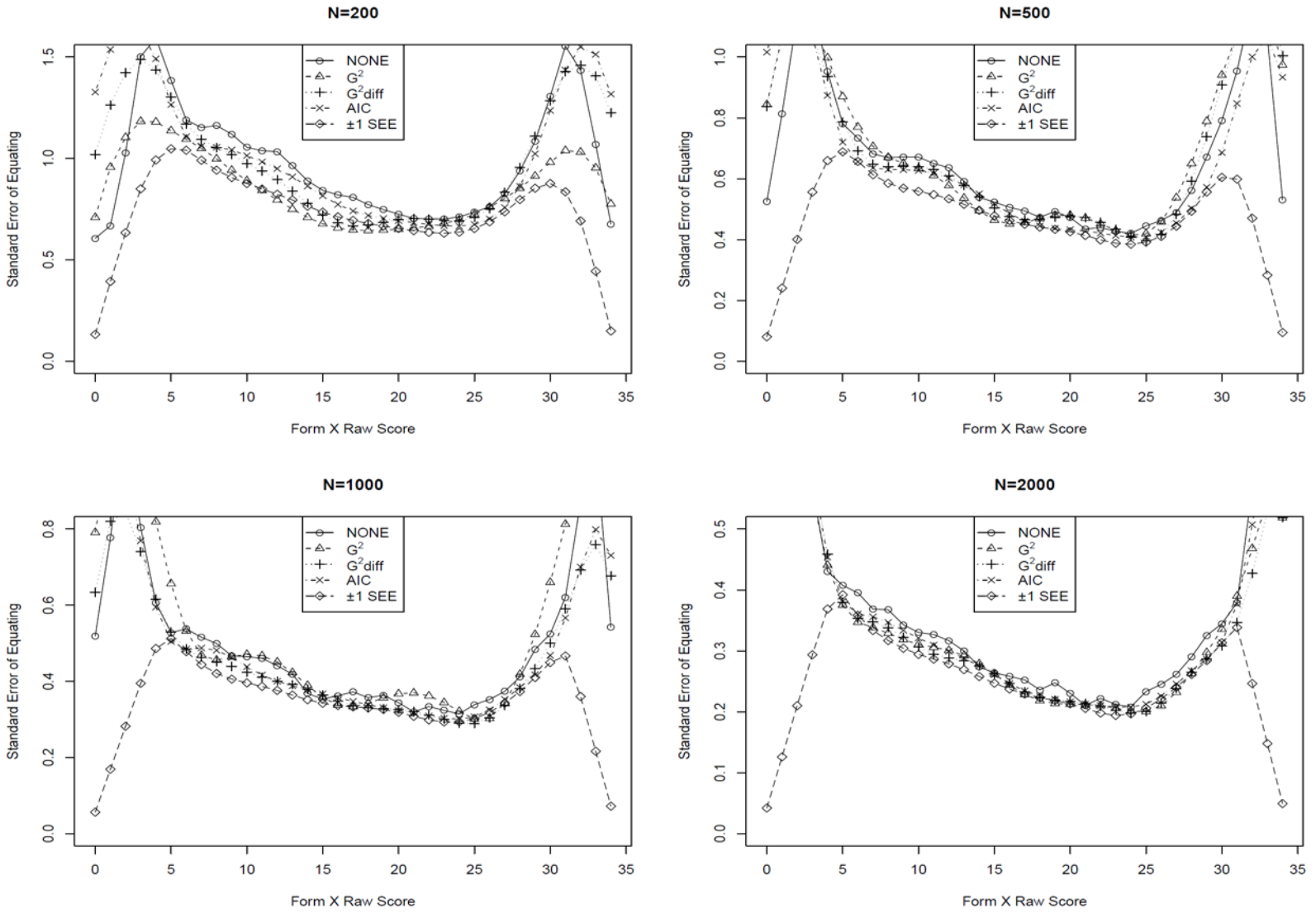

Figure 2. Standard errors of equating for ESS pseudo-test with different smoothing strategies at *N*=200, 500, 1,000, and 2,000.

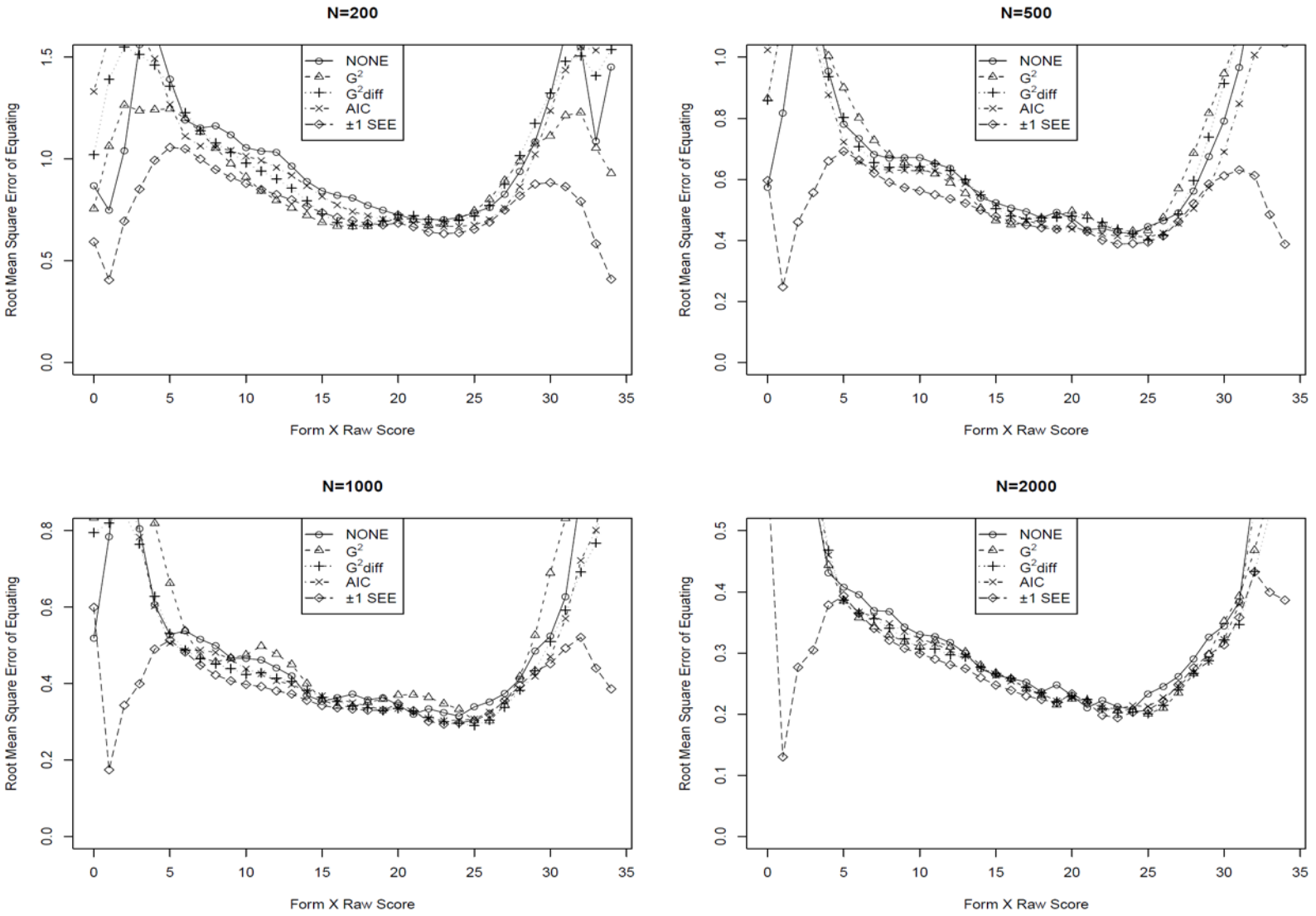

Figure 3. Root mean square errors of equating for ESS pseudo-test with different smoothing strategies at *N*=200, 500, 1,000, and 2,000.