University of Iowa
Iowa Research Online

Theses and Dissertations

Fall 2010

Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions

Sonya Jean Powers
University of Iowa

Copyright 2010 Sonya Powers

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/875

Recommended Citation
Powers, Sonya Jean. "Impact of matched samples equating methods on equating accuracy and the adequacy of equating assumptions." PhD (Doctor of Philosophy) thesis, University of Iowa, 2010.
http://ir.uiowa.edu/etd/875.

Follow this and additional works at: http://ir.uiowa.edu/etd

Part of the Educational Psychology Commons

IMPACT OF MATCHED SAMPLES EQUATING METHODS ON EQUATING ACCURACY AND THE ADEQUACY OF EQUATING ASSUMPTIONS

by

Sonya Jean Powers

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

December 2010

Thesis Supervisor: Professor Michael J. Kolen

ABSTRACT

When test forms are administered to examinee groups that differ in proficiency, equating procedures are used to disentangle group differences from form differences. This dissertation investigates the extent to which equating results are population invariant, the impact of group differences on equating results, the impact of group differences on the degree to which statistical equating assumptions hold, whether matching techniques provide more accurate equating results, and whether matching techniques reduce the extent to which statistical equating assumptions are violated.

Data from one administration of four mixed-format Advanced Placement (AP) Exams were used to create pseudo old and new forms sharing common items. Population invariance analyses were conducted based on levels of examinee parental education using a single group (SG) equating design. Old and new form groups with common item effect sizes (ESs) ranging from 0 to 0.75 were created by sampling examinees based on their level of parental education. Equating was conducted for four common item nonequivalent group (CINEG) design equating methods: frequency estimation, chained equipercentile, item response theory (IRT) true score, and IRT observed score. Groups with ESs greater than zero were matched using matching techniques including exact matching on parental education level and propensity score matching including other background variables. The accuracy of equating results was evaluated by comparing differences between comparison (ES>0) and criterion equating (ES=0) relationships using the root expected mean squared difference statistic, classification consistency for AP grades, and standard errors of equating. The accuracy of equating results and the adequacy of statistical equating assumptions was compared for unmatched and matched samples.

There was relatively little population dependence of equating results, despite large subgroup performance differences. As ES increased, CINEG equating results tended to become less accurate and less consistent. Large differences between criterion and comparison equating relationships appeared to be caused by violations of equating assumptions. As group differences increased, the degree to which frequency estimation and chained equipercentile statistical assumptions held decreased. All exams showed some evidence of multidimensionality. The matching methods that included parental education appeared to improve equating accuracy and the degree to which equating assumptions held for very large ESs.

Abstract Approved: ____________________________________

Thesis Supervisor

____

Title and Department

____

Date

IMPACT OF MATCHED SAMPLES EQUATING METHODS ON EQUATING ACCURACY AND THE ADEQUACY OF EQUATING ASSUMPTIONS

by

Sonya Jean Powers

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

December 2010

Thesis Supervisor: Professor Michael J. Kolen

Copyright by

SONYA JEAN POWERS

2010

All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

____

PH.D. THESIS

____

This is to certify that the Ph.D. thesis of

Sonya Jean Powers

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics) at the
December 2010 graduation.

Thesis Committee: ___________________________________
Michael J. Kolen, Thesis Supervisor

____
Robert Brennan

____
Timothy Ansley

____
Won-Chan Lee

____
Mary Kathryn Cowles

To David, Irma, and Lisa Powers

ACKNOWLEDGMENTS

I am deeply grateful to Dr. Kolen for providing helpful and efficient feedback, without which I would not have been able to complete this dissertation. He is the best advisor anyone could hope to have.

In addition, without the determination of my parents, and the kindness, patience, and encouragement of my sister, I would not have made it to this point.

Many thanks to my committee members who also provided helpful feedback and were very accommodating and encouraging, despite the length of this dissertation. Their classes have been invaluable both for this dissertation, and for my future work in the field.

Finally, I would like to thank Dr. Darrell Sabers who encouraged me to pursue my doctoral degree at the University of Iowa, and who put me in contact with Dr. Kolen. I would not be here were it not for his guidance.

ABSTRACT

When test forms are administered to examinee groups that differ in proficiency, equating procedures are used to disentangle group differences from form differences. This dissertation investigates the extent to which equating results are population invariant, the impact of group differences on equating results, the impact of group differences on the degree to which statistical equating assumptions hold, whether matching techniques provide more accurate equating results, and whether matching techniques reduce the extent to which statistical equating assumptions are violated.

Data from one administration of four mixed-format Advanced Placement (AP) Exams were used to create pseudo old and new forms sharing common items. Population invariance analyses were conducted based on levels of examinee parental education using a single group (SG) equating design. Old and new form groups with common item effect sizes (ESs) ranging from 0 to 0.75 were created by sampling examinees based on their level of parental education. Equating was conducted for four common item nonequivalent group (CINEG) design equating methods: frequency estimation, chained equipercentile, item response theory (IRT) true score, and IRT observed score. Groups with ESs greater than zero were matched using matching techniques including exact matching on parental education level and propensity score matching including other background variables. The accuracy of equating results was evaluated by comparing differences between comparison (ES>0) and criterion equating (ES=0) relationships using the root expected mean squared difference statistic, classification consistency for AP grades, and standard errors of equating. The accuracy of equating results and the adequacy of statistical equating assumptions was compared for unmatched and matched samples.

There was relatively little population dependence of equating results, despite large subgroup performance differences. As ES increased, CINEG equating results tended to become less accurate and less consistent. Large differences between criterion and comparison equating relationships appeared to be caused by violations of equating assumptions. As group differences increased, the degree to which frequency estimation and chained equipercentile statistical assumptions held decreased. All exams showed some evidence of multidimensionality. The matching methods that included parental education appeared to improve equating accuracy and the degree to which equating assumptions held for very large ESs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3PL	Three Parameter Logistic
AP	Advanced Placement
CC	Classification Consistency
CINEG	Common Item Nonequivalent Group
DTM	Difference That Matters
ES	Effect Size
ETS	Educational Testing Service
Form A	New Subform
Form B	Old Subform
Form X	Operational New Form
FR	Free Response
GRM	Graded Response Model
ICC	Item Characteristic Curve
IRT	Item Response Theory
M_0	Matching Method 0: No Matching
M_1	Matching Method 1: Matching on Selection Variable Only
M_2	Matching Method 2: Propensity Score with All Background Variables
M_3	Matching Method 3: Propensity Score without Selection Variable
MAD	Mean Absolute Difference
MC	Multiple Choice
NEAT	Nonequivalent Anchor Test
PEM	Pearson Educational Measurement
PPST	Pre-Professional Skills Test

REMSD	Root Expected Mean Squared Difference
RMSD	Root Mean Squared Difference
SCE	Postsmoothed Chained Equipercentile
SD	Standard Deviation
SE	Standard Error of Equating
SEED	Standard Error of Equating Difference
SEq	Postsmoothed Equipercentile
SFE	Postsmoothed Frequency Estimation
SG	Single Group
UCE	Unsmoothed Chained Equipercentile
UEq	Unsmoothed Equipercentile
UFE	Unsmoothed Frequency Estimation

CHAPTER 1

INTRODUCTION

Equating methods are used to adjust scores on exam forms for differences in difficulty. Forms are often given to different groups of examinees so equating procedures must disentangle group differences from form differences in order to make accurate adjustments. A particularly troubling finding is that various equating methods provide different score adjustments when group differences are large (e.g., Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Stocking & Eignor, 1986). It is unclear which methods, if any, provide accurate adjustments when group differences are large or how large group differences must be before different equating methods provide divergent results. This dissertation investigates the relationship of equating accuracy and group differences. In addition, the impact of group differences on equating assumptions and population invariance of equating results is investigated. Finally, matched samples equating methods, which can be used to make groups more similar, are considered as a possible solution to the equating problems encountered when groups differ substantially in performance from form to form.

In the next two sections, equating designs, methods, assumptions, and properties are described briefly. A more thorough description is provided in Chapter 2. The third section provides a summary of research involving matched samples equating. Finally, the last section of the present chapter provides the goals and research questions that are addressed in this dissertation.

Background

Equating Designs, Methods, and Assumptions

Before one can choose an equating method, the equating design, or data-collection method, must be considered. Kolen and Brennan (2004) describe several equating methods and their respective designs. One of the most common equating designs is the common item, nonequivalent group design (CINEG; also called the nonequivalent anchor test [NEAT] design). This design and the single group (SG) design are used in this dissertation.

In the SG design, one set of examinees takes both (or all) forms of an exam. This design can be employed when there are two or more forms available at one administration and when there is sufficient testing time. Counterbalancing can be done to determine whether order effects have distorted the results. The traditional equating methods (those not based on item-response theory [IRT]) used with the SG design require few assumptions, and are based on fairly straightforward calculations.

Often only one form of an exam is available at a given administration date. Because different examinees take the two (or more) forms, the SG design is not applicable in this situation. A design capable of adjusting for both form and group differences is needed. The CINEG design provides a way to adjust for group differences and form differences by imbedding a subset of items from a previous form into the new form. An assumption of the equating methods used with this design is that performance differences on the common item set indicate group differences that generalize to the rest of the items on the form. The estimate of group differences based on the common items is used so that score adjustments are made based on differences in form difficulty, and not based on differences in group performance.

In order for the common items to provide a good estimate of group performance differences on the exam forms, the common items must be representative of the total exam in terms of content and statistical specifications. Inclusion of a representative common item set becomes challenging when there are passage- or stimulus-based multiple-choice (MC) items, or when there are multiple item types (e.g., MC and free-response [FR]). Operationally, FR items that are time consuming to administer and memorable for examinees are often left out of the common item set. The assumptions of the equating methods used with the CINEG design may be violated when exams with multiple item types, called mixed-format exams, do not contain representative mixed-format common items. The data used in this dissertation are from mixed-format exams. However, the operational common item sets include only MC items. Although MC-only common item sets may not be optimal, the inclusion of FR items in the common item set is outside of the scope of this dissertation.

The CINEG design allows for more flexible testing arrangements than the SG design, but equating methods used with the CINEG design have more stringent statistical assumptions and much more complicated calculations. There are many equating methods that can be used with the CINEG design. Equipercentile equating methods considered in this dissertation include chained and frequency estimation; IRT true and observed score methods are also considered. Equipercentile and IRT methods also are considered for the SG design to parallel the CINEG results. The traditional (non-IRT) SG equating methods require fewer assumptions, and the assumptions may be less likely to be violated. The statistical assumptions and equations for the SG and CINEG equating methods are provided in Chapter 2.

All equating methods tend to provide similar results when the old and new form groups perform similarly (Kolen, 1990). However, when groups differ substantially in performance, the methods can produce different and inaccurate results (e.g., Eignor, Stocking & Cook, 1990; Walker, Allspach, & Liu, 2004). When groups differ, problems with equating results may be caused by violations of equating assumptions. One goal of this dissertation is to investigate the impact of group differences on the adequacy of equating assumptions.

Equating Properties

Ideally any examinee could take any form of an exam, and after the scores are equated to a reference form, the examinee’s equated scores on any and all forms would be identical apart from measurement error. Kolen and Brennan (2004) provide the following five criteria, developed based on the ideal conceptualization of equating results:

1. *Symmetry*-The function used to transform Form X scores to the Form Y scale must be the inverse of the function used to transform Form Y scores to the Form X scale.
2. *Same Specifications*-Forms to be equated must be constructed using the same content and statistical specifications.
3. *Equity*- Examinees with the same true score must have the same distribution of converted scores on all equated forms. Lord's (1980) definition of equity is very stringent. Less stringent definitions include first and second order equity:
   a. *First Order Equity*- Examinees with a given true score have the same mean converted score on all equated forms.
   b. *Second Order Equity*- Examinees with a given true score have the same variance of converted scores, or conditional standard error of measurement, on all equated forms.
4. *Equipercentile Equating Property*- scores on equated forms have the same distribution for a given population of examinees.
5. *Population Invariance*- The equating relationship between forms should be the same for any group of examinees that are a subset of a given population for which the relationship is estimated.

Symmetry and observed score properties are incorporated into most modern equating methods (although observed score properties are not incorporated into the IRT true score equating method). It is also common practice to develop forms for a particular exam based on the content and statistical specifications laid out prior to test development. Although it is possible that exam forms may differ somewhat in difficulty and content, no information was available concerning the adequacy of the test specifications, or the fidelity with which they were applied, for the exams used in this dissertation. However, it is assumed that the test specifications were developed in line with best practice and were properly implemented.

Equity is also an important equating consideration. However, the purpose of this dissertation is to investigate the impact of group differences on equating results. As such, particular attention is given to population invariance in this dissertation.

Matched Samples Equating

Because equating with large group differences may provide inaccurate and inconsistent equating results, a possible solution is to equate with more similar groups. Matched samples equating has been suggested as one possible strategy for eliminating old and new form group differences in performance and/or other characteristics. If matched samples equating reduces the extent to which assumptions are violated or improves the degree to which population invariance holds, then matched sampling may provide better equating results.

Matching individuals in different groups is a common strategy in experimental design where it is used to increase power by decreasing random error. Matching has also been used in quasi-experimental situations when there is no a priori way to ensure that treatment and control groups are randomly equivalent. In situations where participants self-select into treatments, matching the two groups based on the variables that predict group membership may lead to more accurate conclusions about treatment effects.

When there are several variables used to match nonequivalent groups, Rosenbaum and Rubin (1983) propose creating a multivariate composite, which they call a propensity score. A propensity score is the "predicted probability of being in the treatment (versus control) group from a logistic regression equation" (Shadish, Cook, & Campbell, 2002). Once propensity scores are calculated, the treatment and control groups can be matched exactly (if possible), proportionally, based on stratification, weighted, or based on more complicated techniques (see Shadish, Cook, & Campbell, 2002, for a summary of techniques).

The concept of matching nonequivalent groups based on background variables or test scores has also been used in the psychometric field. The nonequivalent groups are not considered control or treatment groups, but rather groups that take different forms (or formats) of an exam. For example, researchers at the Educational Testing Service (ETS) have conducted several studies using common items and background variables to match old and new form groups (see, for example, Dorans, 1990 and the rest of the *Applied Measurement in Education* special issue). These studies did not use propensity scores but did use proportional sampling based on common item scores or student responses to background educational questions. The findings were mixed and are described in detail in Chapter 2.

Other research has looked at matched sampling to compare the scores of examinees that took a computerized form of an exam to those that took a paper-based form of the exam (McClarty, Lin, & Kong, 2009; Way, Davis, & Fitzpatrick, 2006; Way, Lin, & Kong, 2008; Yu, Livingston, Larkin, & Bonett, 2004). In comparability studies, the items are the same but the mode of administration may affect the difficulty of the exam. Typically, examinees are not assigned at random to the computer or paper-based format; they self-select the format, or are non-randomly assigned, for example by school district. Matched sampling has also been investigated as a method for including additional information into the common item set (Moses, Deng, & Zhang, 2009).

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

Research Questions

The main goals of this dissertation are to investigate: 1) the sensitivity of equating methods to group differences, 2) the interaction of group differences, equating assumptions, and population invariance of equating results, and 3) whether or not matched sampling can improve equating results.

Specifically, the research questions this dissertation seeks to address are:

1. To what extent do equating results appear invariant for populations of examinees with different levels of parental education?
2. What is the impact of group differences on equating results?
3. What is the impact of group differences on the degree to which equating assumptions are met?
4. Which matching techniques, if any, provide more accurate equating results?
5. Can matched samples equating reduce the extent to which equating assumptions are violated?

CHAPTER 2

REVIEW OF THE LITERATURE

To ensure score comparability, it is necessary to make statistical adjustments to scores earned on exam forms which differ in difficulty. When the forms are constructed to the same content and statistical specifications, equating methods can be used to adjust the difficulty of exam forms. However, as with any statistical procedure, the accuracy of equating results depends on the extent to which the equating assumptions are met. There are several equating designs, and several applicable equating methods that can be used with each design. Each equating method has its own set of assumptions that may or may not hold in a given situation. Moreover, various methods may be more robust than others to violations of assumptions. The following two sections describe the methods and assumptions of the SG and CINEG designs.

Equating under the SG Design

Traditional Methods

Following the notation in Kolen and Brennan (2004), the equipercentile equating relationship linking Form X scores to the Form Y scale for the SG design is:

$$e_Y(x) = G^{-1}[F(x)], \tag{1}$$

where $F(x)$ is the cumulative distribution function for Form X, and G^{-1} is the inverse of the cumulative distribution function for Form Y. To calculate the equating relationship based on discrete test scores, the percentile ranks of x and y (symbolized $P(x)$ and $Q(y)$ respectively) must be calculated. The following equation is used to find $e_Y(x)$:

$$
\begin{aligned}
e_Y(x) = {} & Q^{-1}[P(x)], \\
& = \frac{\frac{P(x)}{100} - G(y_U^* - 1)}{G(y_U^*) - G(y_U^* - 1)} + (y_U^* - .5), \qquad 0 \le P(x) \le 100, \qquad (2) \\
& = K_y + .5, \qquad\qquad\qquad\qquad P(x) = 100.
\end{aligned}
$$

Here Q^{-1} is the inverse percentile rank function for Form Y, y_U^* is the smallest integer score with a cumulative *100G*(*y*) that is greater than *P*(*x*), and K_y is the number of items on Form Y. Smoothing is often used with equipercentile methods to reduce random error caused by using sample data. For the equipercentile methods used with the SG and CINEG designs, cubic spline postsmoothing was used in this dissertation because it has been found to improve equating results and decrease the standard error of equating (Kolen, 1984; Hanson, Zeng, Colton, 1994).

The assumptions for the SG equating design are:

1. There is a single population *P* of examinees that can take both tests.
2. A random sample from *P* is tested with both Form X and Form Y.
3. The order in which Forms X and Y are administered does not impact performance (i.e., there are no learning or fatigue effects).

The assumption of the equipercentile equating method with the SG design is that the difficulty differences between Form X and Y can be adjusted by setting the distributions equal. The equipercentile method assumption is less restrictive than the linear method in that it could accommodate a linear relationship or a curvilinear relationship between two forms. Linear methods are not considered in this dissertation to keep the number of conditions manageable and because the sample sizes are large enough to use the more flexible equipercentile equating methods.

IRT Methods

Two IRT equating methods used with summed scoring are IRT true score equating and IRT observed score equating. The data used in this dissertation contain both MC and FR items so a dichotomous and a polytomous IRT model must be selected. Because guessing is possible with MC items, the three parameter logistic IRT (3PL) model was chosen as the most likely to fit the data. The mathematical relationship between examinee ability (θ_i) and the probability of a correct response to a given item (p_{ij}) is specified by the 3PL model as:

$$p_{ij} = c_j + (1 - c_j)\frac{exp[Da_j(\theta_i - b_j)]}{1 + exp[Da_j(\theta_i - b_j)]}, \qquad (3)$$

where a_j (item discrimination), b_j (item difficulty), and c_j (pseudo guessing value) are item parameters for item j and θ_i is the ability of person i. D is a scaling constant typically set to 1.7 (Lord, 1980).

For the polytomous items, the logistic graded response model (GRM; Samejima, 1972) was selected. Here the interpretations of D, a_j, b_{jk}, and θ_i are similar to those in the 3PL model, but with the GRM there is a b-parameter for each category k within an item j. p^*_{ijk} is the cumulative category response function.

$$\begin{aligned} p^*_{ijk}\left(\theta_i; a_j, b_2, \dots, b_{jm_j}\right) &= 1, & k &= 1, \\ p^*_{ijk}\left(\theta_i; a_j, b_2, \dots, b_{jm_j}\right) &= \frac{exp[Da_j(\theta_i - b_{jk})]}{1 + exp[Da_j(\theta_i - b_{jk})]}, & k &= 2, \dots, m_j. \end{aligned} \qquad (4)$$

In the SG design, the item responses for both forms are available at the same time. Simultaneous calibration using MULTILOG (Thissen, 1991) ensures that all item and ability parameter estimates are on the same scale, so that a scale transformation is not necessary before conducting equating. This is not always the case with the CINEG design, discussed later.

In IRT true score equating, the true score associated with a given ability level on one form, $\tau_X(\theta)$, is considered equivalent to the true score associated with the same ability level, $\tau_Y(\theta)$, on the other form. True scores on the two forms (also called test characteristic curves) are defined as:

$$\tau_X(\theta_i) = \sum_{j:X} \sum_{k=1}^{m_j} W_{jk} p_{ijk}(\theta_i),$$
$$\tau_Y(\theta_i) = \sum_{j:Y} \sum_{k=1}^{m_j} W_{jk} p_{ijk}(\theta_i), \tag{5}$$

where the summation is over all items in the form and W_{jk} is the integer score associated with item *j* and category *k*. The inclusion of the c_j parameter limits the range of true scores:

$$\sum_{j:X} c_j < \tau_X < K_X,$$
$$\sum_{j:Y} c_j < \tau_Y < K_Y. \tag{6}$$

Therefore the IRT true score equating relationship is defined as:

$$irt_Y(\tau_X) = \tau_Y(\tau_X^{-1}), \tag{7}$$

within the range of true scores specified in equation 6. Solving equation 7 for each true score requires an iterative numeric procedure described in Kolen and Brennan (2004, pp. 177-178).

IRT “true” score equating could be used to equate true scores on Form X to true scores on Form Y if examinee true scores were known. Since they are not, this method is used to convert observed scores on one form to observed scores on another form. Linear interpolation is used to find the equivalents between an observed score of 0 and the sum of the c_j-parameters, and a perfect score on one form is set equal to a perfect score on the other ($K_X = K_Y$).

Rather than apply a true score equating procedure to observed scores, it is possible to estimate the distribution of observed scores on each form using an IRT model, and then use traditional equipercentile equating methods to generate an equating relationship between two forms. IRT observed score equating uses an extension of the Lord-Wingersky recursion formula (Lord & Wingersky, 1984; Thissen, Pommerich, Billeaud, & Williams, 1995) to estimate the conditional distribution of x given θ. The conditional observed score distribution, along with normal quadrature points for the distribution of θ, can be integrated (or summed) across θ to obtain the marginal observed score distributions used in equipercentile equating (see Kolen & Brennan, 2004, pp. 181-185).

The main assumption involved in unidimensional IRT is that responses to all items in the exam measure a single construct. For equating, scores on both 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).

In order to obtain stable estimates of item and person parameters, it is necessary to have large data sets. One complication with the exams that are used in this dissertation is missing data caused by the use of a penalty for guessing. Operationally items were formula scored, with incorrect responses receiving a score of -0.25 or -0.3333 depending on the exam. With formula scoring, students often prefer to skip items they are unsure of rather than chance receiving a deduction in their total score. This makes missing data a large problem. As discussed more extensively in Chapter 3, number correct scoring was simulated in this dissertation by dropping examinees with a large proportion of missing responses, and imputing data for examinees that skipped only a few items. Using imputed data eliminates missing data, but may bias item parameter estimates depending upon how well the imputation reflects the response patterns that would have been obtained had examinees not skipped any items.

A second complication when implementing IRT methods is that the exams considered in this dissertation include "testlets" or a series of MC items that are tied to the same passage, stimulus, diagram, etc. Responses to these items may not be independent. One solution might be to model these items as a polytomous set rather than as discrete dichotomous items. However, for the purposes of this dissertation, unidimensionality is assessed and the degree to which the assumption holds is used to help inform interpretations of IRT equating results.

A mixed-format test design also poses a problem for unidimensional IRT models. Although the MC and FR items are measuring a similar construct, both item types are included under the assumption that FR items measure something that cannot be measured well by MC items. Although it might be better to use a multidimensional model, or estimate a separate unidimensional θ for each item type, multidimensional IRT is beyond the scope of this dissertation. Instead, MC-FR correlations (uncorrected and disattenuated), principal components analysis, polyDIMTEST (Li & Stout, 1995) and polyDETECT (Zhang, 2007) are used to determine whether or not unidimensional IRT equating methods are appropriate for the exams considered in this dissertation.

Equating under the CINEG Design

Traditional Methods

As with the SG design, there are several equating methods that can be used to equate forms with the CINEG design. Unlike the SG equating design where examinees have scores on both/all forms, for the CINEG design, examinees have scores on only one form. Here Form Y is considered the "old form" and Form X the "new form" because the forms are typically administered at different times (one before the other). In order to estimate the performance differences in the two groups, some items are included in both forms. The equating methods used with the CINEG design involve assumptions about the relationship between examinee performance on the common items and performance on the non-common items. Two common equating methods used with the CINEG design are frequency estimation and chained equipercentile.

To define an equating relationship for a single population, the old and new form groups can be weighted to form a synthetic population (s). The following equation defines the curvilinear equating relationship of interest:

$$e_{Y_s}(x) = Q_s^{-1}[P_s(x)], \tag{8}$$

where the synthetic frequencies are calculated as:

$$\begin{aligned} f_s(x) &= w_1 f_1(x) + w_2 \sum_v f_1(x|v) h_2(v), \\ g_s(y) &= w_1 \sum_v g_2(y|v) h_1(v) + w_2 g_2(y). \end{aligned} \tag{9}$$

Here $f_s(x)$ represents the distribution of scores on Form X in the synthetic population, $g_s(y)$ represents the distribution of scores on Form Y in the synthetic population, $f_1(x)$ represents the distribution of scores on Form X in Population 1 (the population who took Form X), $g_2(y)$ represents the distribution of scores on Form Y in Population 2 (the population who took Form Y), $f_1(x|v)$ represents the conditional distribution of Form X scores given common item (v) scores in Population 1, $g_2(y|v)$ represents the conditional distribution of Form Y scores given v in Population 2, $h_1(v)$ and $h_2(v)$ represent the common item score distributions in Populations 1 and 2 respectively, and w_1 and w_2 represent the weights for Populations 1 and 2 respectively. The details of calculating the unobserved quantities are provided in Kolen and Brennan (2004). Estimation of the quantities for the frequency estimation method is based on the assumption that the conditional distribution of total score (X or Y) given the common item score (V) is the same in both populations.

For the chained equipercentile method, the equipercentile relationship between X and V is calculated using data from the new form group (symbolized $e_{V1}(x)$). Then the equipercentile relationship between V and Y is calculated using data from the old form group (symbolized $e_{Y2}(v)$). Finally, the equipercentile relationship between X and Y is calculated by chaining the two previous results ($e_{Y(chain)} = e_{Y2}[e_{V1}(x)]$). The assumptions for the chained method are (von Davier, Holland, & Thayer, 2004a):

1. For a given population, the link from X to V is group invariant.
2. For a given population, the link from V to Y is group invariant.

IRT Methods

IRT true score and observed score equating methods are used with the CINEG design in much the same way they are used with the SG design. Often item calibration is done separately for old and new forms with the CINEG design because the forms are administered at different times. With separate item calibrations, item (j) and ability (i) estimates are on separate scales (I and J). An additional scale transformation step is needed to place item and ability estimates on the same scale. The scale transformation method used in this dissertation was the Haebara (1980) method. The Haebara method finds the transformation constants that minimize the criterion function (*Hcrit*), where *Hdiff* sums over the set of common items (V), and *Hcrit* cumulates *Hdiff* over examinees (i):

$$Hdiff(\theta_i) = \sum_{j:V} \sum_{k:j} \left[\begin{array}{c} p_{ijk}\left(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj2} \ldots \hat{b}_{Jjk}, \ldots \hat{b}_{Jjm_j}, \hat{c}_{Jj}\right) - \\ p_{ijk}\left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij2} + B, \ldots, A\hat{b}_{Ijk} + B, \ldots A\hat{b}_{Ijm_j} + B, \hat{c}_{Ij}\right) \end{array} \right]^2, \tag{10}$$

$$Hcrit = \sum_i Hdiff(\theta_i).$$

After calculating the transformation constants based on the common items, the transformation is applied to all items. When all item parameter and ability estimates are on the same scale, IRT true and observed score equating is conducted. *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009), a set of open-source C functions for equating, was used to implement the scale transformation and equating procedures used in this dissertation.

The IRT parameter estimation and equating assumptions are similar for the CINEG and the SG design. However, for IRT observed score equating, there are four observed score distributions estimated for the CINEG design but only two for the SG design. The CINEG design has the additional assumption that the same construct is measured in the two groups of examinees and on both forms. Additionally, because there are two groups with the CINEG design, weights must be considered for specifying the synthetic population for IRT observed score equating.

Population Invariance

One desirable equating property is that the equating relationship should be the same for any subgroup of examinees from the same population. 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. Dorans and Holland (2000) defined two statistics to quantify the difference between linking functions computed for the whole group and linking functions computed for each of the subgroups. These statistics were developed for the SG or random group equating design and von Davier, Holland, and Thayer (2004a) demonstrated their applicability to the CINEG design. The first statistic is a root mean squared difference statistic (RMSD) that quantifies differences in the equating relationship, at a given Form X score, in terms of the subgroup relationship ($eq_{Y_h}(x)$) and the full group equating relationship ($eq_Y(x)$). The weight (w_h) is the proportion of examinees from the whole group contained in subgroup h:

$$RMSD(x) = \frac{\sqrt{\sum_{h=1}^{H} w_h [eq_{Yh}(x) - eq_Y(x)]^2}}{\sigma(Y)}, \qquad (11)$$

$$w_h = N_h/N.$$

The denominator of the RMSD statistic is the standard deviation of the Form Y scores which allows for the interpretation of RMSD(x) in terms of Form Y standard deviation units. Dorans and Holland's (2000) Root Expected Mean Standardized Difference (REMSD) statistic provides a summary statistic that quantifies equating relationship differences across the entire score scale:

$$REMSD = \frac{\sqrt{\sum_{h=1}^{H} w_h \sum_{\min(x)}^{\max(x)} v_{xh} [eq_{Yh}(x) - eq_Y(x)]^2}}{\sigma(Y)}, \qquad (12)$$

$$v_{xh} = N_{xh}/N.$$

Here the conditional equating difference ($eq_{Yh}(x) - eq_Y(x)$) is squared and then multiplied by the proportion of examinees at a particular Form X score (v_{xh}). The result is summed across the Form X scale (from min(x) to max(x)), and finally, weighted again by subgroup size (w_h).

Several other population invariance statistics have been suggested (Kim, 2006; Kolen & Brennan, 2004). These statistics change the way in which subgroups are weighted and provide pairwise comparisons of subgroups rather than an index of all subgroup differences referenced to the full group equating relationship. Interpretation of RMSD and REMSD statistics has often centered around "the difference that matters" (DTM) criterion. The DTM is defined as half of a reported score unit which is a difference that may result in a different score for examinees. When RMSD or REMSD statistics exceed the DTM (standardized to be on the same scale), the differences between equating results is considered to be of practical significance (Dorans, 2003). However, differences between subgroup equating relationships might also be compared in terms of the standard error of equating (SE). Because subgroup sample sizes are much smaller than the full group sample size, the difference between subgroup equating relationships might be a result of random sampling error rather than a true difference. Therefore bootstrap SEs are considered where possible in this dissertation when interpreting the results of population invariance analyses.

ETS conducted four separate studies of population invariance for 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.

Using data from the 2001 and 2003 administrations of AP Calculus AB, von Davier and Wilson (2006) investigated the population invariance of IRT true score equating. The subpopulations considered were males and females. First, population invariance was evaluated with MC items only; subsequently, MC and FR items were both included in the analyses although in both cases the common items were only MC items. For IRT true score, Tucker, and chained equipercentile equating, very little population sensitivity was found even when the FR items were included in the linking design. The authors suggested that the high correlation between MC and FR items might explain this finding. However, RMSD values tended to show population sensitivity towards the extremes of the distribution. The weighting process used in the REMSD statistic hid these differences. The authors noted that Dorans, Holland, Thayer and Tateneni (2003) reported population sensitivity when linking Calculus AB from 1999 to 1998 but little to no population sensitivity for 2000 to 1999. They tentatively concluded that population invariance might not be a stable characteristic of examinations across administrations and suggested that similar studies be carried out over several administrations.

These studies used data similar to that used in this dissertation. In general they provide limited information about when population dependent equating results might be anticipated. However, they provide a framework for investigating the equating property of population invariance. Furthermore, the studies cited above use many of the same methods and statistical criteria that are used in this dissertation. A goal of this dissertation is to compare the impact of subgroup performance differences (population invariance) and administration group performance differences on equating accuracy.

Comparison of Equating Methods

Four nonlinear equating methods are considered for the CINEG design in this dissertation: frequency estimation, chained equipercentile, IRT true score, and IRT observed score. If a “true” equating relationship exists, it would seem desirable for all four methods to produce the same “true” relationships. However, equivalent results are unlikely to be obtained in practice because each method has its own underlying statistical assumptions, which hold to varying extents.

According to Kolen (1990), when group differences are fairly small and exam forms and common items are constructed to be nearly parallel in terms of content and statistical properties, all equating methods tend to give reasonable and similar results. Many empirical studies provide evidence in support of Kolen's conclusion (e.g., Cook, Dunbar, & Eignor, 1981; Cook, Eignor, & Taft, 1998; Marco, Petersen, & Stewart, 1979).

For the chained equipercentile and frequency estimation methods, von Davier, Holland, and Thayer (2003, 2004) provided a mathematical proof of the equivalence of the two methods under two extreme cases. First, given equivalent old and new form groups, that is, equal common item score distributions in both groups, the results produced by either the chained equipercentile or the frequency estimation method are equivalent. Secondly, given a perfect correlation between the common item scores and the full form scores, both chained equipercentile and frequency estimation provide identical solutions. Although these two cases are extreme, they provide some indication that chained equipercentile and frequency estimation methods should provide reasonably similar results with when group differences are small and when common item scores are highly correlated with total test scores.

What to expect from the two IRT methods is less clear in part because the psychometric models used with IRT have very different assumptions from those involved with the traditional chained equipercentile and frequency estimation methods. In addition, one IRT methods is an observed score equating method, while the other IRT method is, at least in theory, a true score equating method.

When group differences are large or when common item-total test correlations are small, how different are equating results likely to be, and which equating methods produce more accurate results?

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.

Han, Kolen, and Pohlmann (1997) also compared the equating results of equipercentile, IRT observed, and IRT true score methods, using the equating a test to itself criterion. The authors found that IRT true score equating was closest to the criterion, followed by IRT observed score, and then equipercentile. However, 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. In addition, Lord and Wingersky (1984) found that IRT true and observed score methods produced very similar results when equating a test to itself through an equating chain involving six equatings. Though neither method provided the identity equating relationship across the entire score scale, both methods were very close. These studies suggest that equating accuracy is similar for both IRT methods.

Harris and Kolen (1986) compared the linear, equipercentile, and IRT true score methods with the random groups design. The criterion in this study was the population invariance of equating results when comparing the equating relationships for groups constructed to be high and low in performance. All methods provided fairly population-invariant results. The authors concluded that population-invariance was not a useful criterion for preferring one method over another.

Using the CINEG design, 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 was only 0.07. All three equating methods produced equating relationships that were within the DTM in terms of REMSD. Only the linear method exceeded the DTM at high score points in terms of RMSD when FR items were not included in the equating. Inclusion of FR items resulted in higher RMSD values for all methods. This finding is especially relevant for this dissertation because of the mixed-format data involved.

With nonequivalent groups and common items that are not a perfect measure of total test scores (as is always the case with operational data), the results for frequency estimation and chained equipercentile may differ to some extent. Several studies have compared equating results for frequency estimation and chained equipercentile. Harris and Kolen (1990) compared the two methods for groups that differed by less than 0.1 standard deviations, and groups that differed by more than a third of a standard deviation. The authors found that the two methods differed by more than the frequency estimation method SE. 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). The frequency estimation method appears to have slightly less random equating error compared to the chain equipercentile method (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008). However, 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. In fact, the longer the relative length of the common item set, the less bias found for both methods (Holland, Sinharay, von Davier & Han, 2008; Wang, Lee, Brennan, & Kolen, 2008). This is not surprising given that as the length of the common item set increases, so does the common item-total test score correlation. And, as proved by von

Davier, Holland and Thayer (2003, 2004), when the correlation reaches one, the results of the two methods converge.

In a study by von Davier, Holland, and Thayer (2003), the equating results produced by 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 (Eignor, Stocking, & Cook, 1990; Cook, Eignor, & Schmitt, 1990). Another study found that the sensitivity of equating methods to group differences appears to differ across exams (Livingston, Dorans, & Wright, 1990). Finally, Wright and Dorans (1993) suggested that the type of sampling or matching methods used determines the stability and accuracy of the equating results for different methods. Based on these results, it appears that the degree to which population invariance holds for a given equating method depends on multiple variables, some of which may not be easily controlled in operational testing situations. Even if one method provides more population invariant results in a particular situation, it may not provide optimal results in another situation.

Matched Samples Equating

In some research situations, additional information is known about the study participants. In an experiment where subjects are assigned to conditions, matching two or more treatment/control groups on a variable or set of variables related to the outcome can improve the power of the test for group differences. Matching can also be useful in quasi-experimental designs where the researcher does not have control over the assignment of individuals to treatments. Like ANCOVA, matching cannot completely remove the effects of nonrandom assignment to conditions. However, both methods can be used after the data have been collected to help disentangle the true treatment effect from pre-existing group differences caused by nonrandom self-selection of participants into conditions.

Ideally groups would be matched on the selection variable or variables—those characteristics that cause individuals to select one treatment condition over another. In practice, however, selection variables are not known, so any demographic or background variables that might impact treatment membership, and that are available to the researchers, are used to match. According to Shadish, Cook, and Campbell (2002):

> The goal is to include all variables that play a role in the selection process (including interactions and other nonlinear terms; Rosenbaum & Rubin, 1984; Rubin & Thomas, 1996) and that are presumptively related to outcome, even if weakly so. (p. 162)

With large numbers of background variables with which to match, the number of participants with any given combination of values tends to be very small. Rosenbaum and Rubin (1983, 1985) suggest using a multivariate composite of background variables they call a propensity score. The propensity score can be defined as "the conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983). Conditional probabilities are calculated using logistic regression (Moses, Deng, & Zhang, 2009; Pampel, 2000).

$$Propensity(P|\boldsymbol{X}) = \frac{1}{1 + e^{-b_0 - b_1 x_1 - b_2 x_2 - \cdots - b_n x_n}} \qquad (13)$$

Here *P* stands for membership in the reference treatment, which could be the control group, ***X*** represents the vector of observed covariates, and b_0 through b_n are regression coefficients.

Once propensity scores are calculated, a variety of matching methods can be used. Rosenbaum and Rubin (1985) described three matching techniques that might be considered:

1. Nearest available matching on the estimated propensity score.

This method is the computationally simplest of the three methods. Propensity scores are calculated based on membership in a reference treatment. Regardless of actual group membership, individuals areassigned estimated propensity scores based on their specific vectors of background variables. Finally, reference and comparison group individuals are matched pairwise based on their estimated propensity scores. If, for example, the control group was the reference group, and an individual in the control group had a propensity score of 0.78, then the individual in the treatment group with the propensity score closest to 0.78 would be matched with the control group individual. The two participants would be removed from the pool of unmatched control and treatment individuals, and the process would be repeated until as many of the control group individuals had matches as possible.

In practice, some rule must be established for balancing the closeness of matches with the proportion of control group individuals that is matched. The closer the matches must be, the less likely that all control group individuals have a match. Parsons (2001) provided a SAS macro to match individuals on 5 to 1 decimal places. The process begins by matching those individuals whose propensity scores are the same to within five decimal places of precision. When no more matches are found, the remaining pool of unmatched individuals are matched based on four decimal places of precision. The iterations continue until all individuals that have the same propensity scores within one decimal place have been matched. Based on this procedure, Parsons was able to match 85% of reference group members while simultaneously maximizing the closeness of the matched pairs.

2. Mahalanobis metric matching including the propensity score.

This matching method is the most computationally burdensome. According to D'Agostino (1998):

> Mahalanobis metric matching is employed by randomly ordering subjects, and then calculating the distance between the first treated subject and all controls, where the distance, d(i,j) between a treated subject i and a control subject j is defined by the Mahalanobis distance… The control subject j, with the minimum distance d(i,j) is chosen as the match for the treated subject i, and both subjects are removed from the pool. This process is repeated until matches are found for all treated subjects. One of the drawbacks of this technique is that it is difficult to find close matches when there are many covariates included in the model. (p. 2268)

Mahalanobis distance is defined as:

$$d(i,j) = (u - v)^T C^{-1} (u - v) \tag{14}$$

where *u* and *v* are values of the matching variables for treated subject *i* and control subject *j*, and *C* is the sample covariance matrix of the matching variables from the full set of control subjects (D'Agostino, 1998). The propensity score is included in the calculation of *d*(*i*,*j*) in this technique.

3. Mahalanobis metric matching within calipers defined by the propensity score.

Rosenbaum and Rubin (1985) suggest that this method is preferable to the other two. It is also less computationally demanding than matching procedure two. Here, propensity scores are calculated for all participants in both groups. Then for a given reference group participant's propensity score, any participants in the other group who have a propensity score within "calipers" (typically 0.25 standard deviations of the target propensity score), are considered as potential matches. The Mahalanobis distance is then calculated for all potential matches within the calipers, and the participant with the smallest distance is paired with the reference group individual, and removed from the pool. This process is repeated until as many matches are found as possible.

Many other more or less complicated matching techniques have also been considered. Shadish, Cook, and Campbell (2002) mention exact matching (essentially

Rosenbaum and Rubin’s Method 1), caliper matching (essentially Rosenbaum and Rubin’s Method 3), index matching, cluster group matching, benchmark group matching, optimal matching, and stable matched bracketing.

Propensity score matching is useful in that it reduces the number of matching variables to one, and because it “only involves covariates and not outcome variables, there is no chance of biasing results in favour of one treatment condition versus the other through the selection of matched controls” (D’Agostino, 1998). The problem is that the set of observable covariates cannot completely eliminate the effect of nonrandom assignment of individuals to treatments. Moreover, matching may not actually help to eliminate treatment selection effects or bias. When non-redundant predictors of group membership are not included in the model, the treatment groups are undermatched. Matching can actually increase the bias over not matching in some cases, leading to invalid conclusions about the effectiveness of treatments (Shadish, Cook, & Campbell, 2002).

Nonequivalent groups in an equating context parallel nonequivalent groups in quasi-experimental situations. The groups self-select into administrations based on a set of unknown variables, confounding group differences and form (“treatment”) differences.

Researchers at ETS have been investigating the usefulness of matching nonequivalent groups for equating purposes since the late 1980s. Cook, Eignor, and Schmitt (1988), attempted to adjust for group differences in spring old form and fall new form groups. The equating results using a spring old form were different from the equating results using a fall old form because the spring group performance was much higher than the fall group performance. Two matching strategies were considered: matching the old and new form common item score distributions, and matching using two self-reported background questions related to science ability. The exam considered in this study was a Biology achievement test. Because the fall groups did not differ greatly from year-to-year, the fall-to-fall equating results may be closer to providing more accurate equating results. Based on this logic, the fall-to-fall equating relationship was used as the criterion equating relationship. Equating relationships for spring-to-fall unmatched, matched based on common item score distributions, and matched based on background questions, were compared to the fall-to-fall criterion relationship.

The authors also compared the results of equating with Levine, Tucker, equipercentile, and IRT equating methods. Matching appeared to result in more agreement among the traditional equating methods. However, all equating results (matched and unmatched) underestimated the criterion scaled score mean and standard deviation, although matching on common items appeared to bring the results closer to the criterion for all methods except the Tucker method. Matching on the background questions did not appear effective.

In a special issue of *Applied Measurement in Education* (volume 3, issue 1), ETS researchers presented the results of four additional matched samples equating studies. Dorans (1990) provided an introduction to the issue where he described equating designs, equating methods, and sampling methods. Three sampling methods were used in the following four studies.

1. *Representative Sampling.* With this sampling method, the old form and new form are equated as usual. No matching is done.
2. *New Form Matched Sampling*. With this sampling method, the old form is matched to the new form in terms of the scores on the common item set.
3. *Reference Population Sampling.* Similar to the conceptualization of a synthetic population, this method matches both the old form and the new form common item score distributions to a target distribution.

Dorans (1990) also described proportional matching, used when exact matching cannot be achieved:

> Proportional matching for new-form matched sampling proceeds as follows. First, relative frequencies on the anchor are computed in the new-form sample and the old-form sample. Then,

> the ratios of the old-form and new-form sample relative frequencies are computed for each score level. Ratios above 1 indicate that exact matching is possible at that score level, and ratios below 1 indicate that exact matching is impossible. The proportional matching coefficient is chosen from among the ratios below 1….Approximate proportional matching involves borrowing examinees from adjacent cells to fill up cells whose ratios fall below the proportional matching coefficient. (Dorans, 1990, pp. 15-16).

The first study (Lawrence & Dorans, 1990, also reported as Lawrence & Dorans, 1988) investigated the effect of matched samples equating using the common item set as the matching variable for the SAT Math and Verbal sections. Equating methods used in this study included Tucker, Levine observed score, chained and frequency estimation equipercentile, and IRT true score. The old form and new form groups differed in performance on the common items by as much as 0.39 standard deviation units. Equating results were not compared to a criterion but judged based on consistency of results. In general, the equating methods provided more consistent results when using matched sampling. The Tucker and frequency estimation methods did not appear to differ much for the representative and matched samples. The equating results for other equating methods did change when the sampling procedure changed.

The second study, by Eignor, Stocking, and Cook (1990), provided a follow-up of the Stocking, Eignor, and Cook study (1988). As the 1990 paper provides all of the results and conclusions from the 1988 paper, only the 1990 paper is summarized here. Simulation using IRT was used to determine whether or not matched samples equating provided accurate results. Item parameter estimates for an SAT Verbal test were considered the true values. Because formula scoring was used with this test, the authors also simulated a missing data condition. Equating methods compared included Tucker, Levine observed score, chained equipercentile, and IRT true score. The criterion in this study was based on equating a test to itself with no examinee differences based on the 3PL IRT model. Results indicated that the Tucker method was least sensitive to sampling method, but farthest away from the criterion equating relationship. The other methods were more affected by sampling method, and had equating results further from the criterion equating relationship when matched sampling was used. The authors suggested that matching on common items might not be advisable.

Schmitt, Cook, Dorans, and Eignor (1990) looked at the impact of matched sampling under similar conditions to those used in Cook, Eignor, and Schmitt (1988). The same exam was used—namely the College Board Biology Achievement Test. The criterion was again the fall-to-fall equating relationship (the Tucker method was used operationally and serves as the criterion method for Schmitt et al.). The equating methods compared included Tucker, Levine true score, chained and frequency estimation equipercentile, and IRT true score. Matching was done based on the common items to adjust for performance differences between spring and fall administration groups. The authors found that matching resulted in more consistent equating results across methods, and matched equating results that were closer to the criterion equating relationship. The Tucker and frequency estimation methods were less affected by sampling design compared to the other methods. IRT true score equating was most affected by sampling design.

The last study by Livingston, Dorans, and Wright (1990) used SAT Math and Verbal scores to simulate group differences, and then used matched sampling based on common items to adjust for the differences. Specifically, the old form group was made less able for the verbal test by selecting an examinee group approximately a third of a standard deviation below the mean on the math test. A similar procedure, using verbal scores, was used to create group differences on the math test. The criterion equating in this study was the equipercentile equating relationship obtained from a random groups design with 115,000 students taking each form. The equating methods compared in this study included Tucker, Levine observed score, frequency and chained equipercentile, and IRT true score. Again, the equating results tended to converge when matched sampling was used, however the results did not accurately reflect the criterion equating relationship. The authors concluded that matching on an anchor test created a regression effect and biased results. Matching on the observed common items scores did not properly adjust for true group differences because of regression effects.

In the first of two reviews in the *Applied Measurement in Education* special issue, Kolen (1990) provided a summary and comments based on the ETS research. He noted that the consistency of the Tucker method when there are extreme group differences has not been found in other studies (e.g., Petersen, Cook, & Stocking, 1983; Petersen, Marco, & Stewart, 1982). Kolen also briefly reviewed the literature comparing traditional and IRT methods. In general, research has shown that all equating methods appear to work reasonably well and provide similar results for the CINEG design when the groups are fairly similar and forms and common items are well constructed. However, as the performance of old and new form groups becomes more dissimilar, the equating methods produce divergent and inaccurate results. Matching on common items appears to be a straightforward way to create groups of similar performance for establishing an equating relationship.

However, based on the results presented in the special issue, Kolen (1990) concluded that matching on common item scores did not provide good equating results. For the Tucker and frequency estimation methods, the results were often similar whether matching was used or not. For the other equating methods the results were often less biased when using unmatched data. Kolen suggested that other matching variables be considered especially those related to examinees educational experiences. However he also warned that “if the common items do not behave in the same way in the old and new groups, then no equating method can be expected to function adequately” (Kolen, 1990, p. 100).

The second reviewer, Skaggs (1990), also provided a very thorough comparison of the findings across the methods. One important limitation of the matched samples equating studies was that they did not provide SEs, so it was difficult to tell which differences between equating methods and sampling procedures were relevant, and which were within sampling error. Skaggs concluded: “The answer to the question of matching or not matching resides with each test and method. An answer cannot be determined a priori from these articles” (p. 111).

Two additional follow-up studies of matched samples equating were conducted by ETS researchers. Cook, Eignor, and Schmitt (1990) followed up their 1988 study by looking at other tests including College Board Chemistry, Mathematics Level II, American History and Social Studies, and French. The same equating methods and criteria were used as in 1988. They also added another matching variable for the French test based on student responses to a background question. In contrast to the previous studies, the Tucker and frequency estimation methods appeared more sensitive to group differences and the Levine and chained methods appeared less sensitive. Consistent with other studies, matching on common items resulted in very consistent equating results across all equating methods. Unfortunately the matched samples equating results always resulted in over- or under-estimated scale score means when compared to the criterion equating relationship. Matching on background questions was considered separately rather than in conjunction with common items. Unlike matching on the common items, matching on the background questions did not result in consistent equating relationships nor did it bring the equating results closer to the criterion relationship.

The second study by Wright and Dorans (1993) provided a follow-up to Livingston, Dorans, and Wright (1990). Recall in the original study that group differences between old and new groups were created by selecting examinees based on their other SAT score. For example, if equating was going to be conducted for the math test, verbal scores were used to select lower performing examinees for one of the forms. In the 1993 study the authors attempted to rematch examinees based on the scores from the other SAT exam. This method may have been more successful than simply matching on the anchor test because it is similar to a small random sample from the original populations, which were very similar. The authors also tried using the scores from the other test as common items and found that this method worked fairly well for Tucker and frequency estimation methods, but not well for Levine and chained equipercentile methods. This study indicated that matching on the selection variable (i.e., the variable that causes group differences) improves equating results. Unfortunately, unless the data are manipulated by the researcher, it is impossible to know which variables distinguish group membership. Propensity scores can be used to predict group membership based on any covariates that researchers have on examinees, but the variables available may not be the underlying selection variables of interest.

Paek, Liu, and Oh (2006) used propensity score matching to develop a linking relationship between the PSAT and the SAT. The PSAT is a shorter exam given to examinees in 10th and 11th grades whereas the SAT is a longer exam given to examinees in 11th and 12th grades. SAT examinee performance was over a third of a standard deviation higher than PSAT examinee performance on the common items. The authors compared four equating conditions: random-groups-no matching, CINEG-no matching, random-groups-matching (without common items), and random-groups-matching (with common items). With the random groups design, direct linear and equipercentile equating methods were used. With the CINEG design, Tucker and frequency estimation methods were used.

Matching variables included sex, ethnicity, and high school grade level, plus common item scores for the matching with common items condition. The propensity score distributions for the combined PSAT and SAT groups were divided into 20 strata based on their percentile ranks. Examinees within the same strata were matched. The authors found that none of the matching conditions provided more accurate results than the CINEG-no matching condition. Matching on the common items provided results closest to the CINEG-no matching condition. Matching without common items did not provide good equating results. The authors suggested that the low predictive power of the matching variables they used (sex, ethnicity, and grade) for explaining group differences, compared to common item scores, accounted for the results.

Despite the pessimistic outlook on matched samples equating from the ETS research reviewed above, matched sampling has also been studied more recently by ETS and Pearson Educational Measurement (PEM) as a way to study mode effects when forms are administered in paper and computerized formats. In these studies, the items taken by both groups are the same but the groups self-select into administration mode and differ in ability and a variety of other characteristics. Although the items are the same, there is concern that administration mode may affect item difficulty because of issues like speededness, computer skills, and differences in the way items are displayed. In order to draw conclusions about mode effects, it is necessary to eliminate the confounding influence of group differences. Matching samples on background variables has been investigated as a possible solution to eliminate the confounding of group and mode differences.

In an ETS study by Yu, Livingston, Larkin, and Bonett (2004) matched sampling was used with the Pre-Professional Skills Test (PPST) to try to compare a paper-based administration group to a computerized administration group that differed by approximately half a standard deviation. The authors used logistic regression to assign propensity scores to examinees. Variables in the model included gender, ethnicity, educational background, job related information, and teaching experience. The propensity score distribution was divided into 20 intervals. The examinees that took the test on the computer were matched to the examinees that took the test on paper by assigning a weight to each examinee based on the propensity score interval they were within. The mean performance difference between the two groups remained at approximately 0.5 standard deviations. The essay items for the computerized version of the PPST are not always the same as those given to the paper group. However, with matched sampling using only data where the two groups had the same essays, the effect size difference remained. Remaining group differences could indicate either that the matching procedure did not control adequately for administration selection effects or that there was an administration mode effect.

Researchers at PEM have used matched sampling to compare mode effects and to use different score conversion tables for computer and paper groups should the equating relationships for the two groups be extremely discrepant. For example, Way, Davis, and Fitzpatrick (2006) used matched sampling to:

> …evaluate the comparability of online and paper versions of the Texas Assessment of Knowledge and Skills (TAKS) in mathematics, reading/English language arts, science, and social studies at grades 8 and 11 for the purpose of test score reporting, and to appropriately adjust equated score conversion tables for students testing online as warranted. (p. 3)

The matched sampling method used in the Way, Davis, and Fitzpatrick study was much different than the study by Yu et al. (2004). Instead of using propensity scores based on background variables, the authors matched the larger paper administration group to the computer administration group based on examinees previous years' scores on a paper and pencil version of TAKS. Exact matching for the 8th grade data was done on math and reading scores from grade 7. In grade 11, no matching was done because examinees were randomly assigned to the administration modes. After matching (or random assignment), the computer scores were equated to the paper scores based on the random groups equating design. IRT parameter estimation was conducted using the Rasch model. The equating process was bootstrapped 500 times to estimate the SE. Finally, the average equating results were compared to the identity equating relationship (i.e., no equating), and were considered the same within sampling error if the equating relationships were within two SEs. If mode effects were not present, then the equating relationship for the random group design is equivalent to equating a test to itself and should be identical to the identity relationship within random error. Differences beyond two SEs were considered especially important around the "meets standard" cut point.

The authors found that the computer version of TAKS was more difficult than the paper version for both grade levels and both content areas. Because of these differences, and the high stakes decisions made based on the scores (passing the 11th grade TAKS is a graduation requirement), separate score conversions were developed for the online and paper administration groups for reporting purposes.

Way, Davis, and Fitzpatrick (2006) also conducted a simulation study to assess the sensitivity of the matched samples comparability analysis (MSCA) procedure given different magnitudes of mode effects. They found that the MSCA procedure did not always identify form differences of only 0.25 raw score points but that mode effects as large as 1 raw score point were always identified.

Way, Lin, and Kong (2008) provide a summary of results for 46 studies in 5 different states. In all matching studies, previous test scores were used to match, but the authors also described using background variables if they are available. In a majority of cases, even when matching was not performed because random assignment was possible, different score conversions were developed for the two administration groups.

Another sensitivity analysis was conducted by McClarty, Lin, and Kong (2009) using simulated data under a variety of conditions including sample size of the paper administration group (60000, 20000, 5000), sample size of the computer administration group (1500, 1000, 500, 250, 100), mode effect (0, 1, and 2 raw score points), and IRT model (Rasch, 3PL). Item parameters were generated using the 3PL model for the test of interest and for two matching tests from prior years. The authors found that their simulated groups could be matched exactly even with small sample sizes. In addition, no mode effects were found for the 0 raw score point mode effect. For the 1 and 2 raw score points, sample sizes of 1000 or 1500 were necessary to detect the score difference consistently with the Rasch and 3PL models. Score differences were more frequently identified with the Rasch model than with the 3PL model because of the smaller SE for the Rasch model.

Summary

This chapter described the equating methods used with the SG and CINEG designs. Equipercentile methods with cubic spline postsmoothing and IRT equating methods were considered along with the assumptions underlying each method. Methods used with the SG design require far fewer statistical assumptions than methods used with the CINEG design. The IRT equating methods involve the most stringent statistical assumptions.

Research and statistics developed for the evaluation of population invariance were described. Also, matching methods were described both as they have been developed in experimental design, and as they have been adopted for use in equating. Though the research conducted at ETS has not found matched samples equating particularly promising, the researchers did not use propensity scores and primarily used common item scores as the matching variable. Matching on the common items biased results but appeared to bring many of the equating results closer to the criterion equating relationship and made them more consistent with each other. Using other examinee background variables may prove more effective.

Matched sampling is also used in practice by PEM to determine whether or not administration mode effects are large enough to require separate conversions for paper and computer administration groups. If matched sampling is not an effective way of controlling for group differences, then use of matched sampling operationally may be questionable.

The research reviewed in this chapter provides a brief overview of equating assumptions, population invariance of equating results, and sampling procedures. However the research in these areas has not been well integrated. The goal of this dissertation is to investigate equating assumptions, properties, and sampling procedures in order to better understand how they interact and to suggest how these findings might improve equating practice.

CHAPTER 3

METHODS

In this chapter, the set of methods used to address the five research questions is provided. The first section identifies the source of the data, the exam forms and scores used in data analysis, and the construction of subforms A and B, which were used in all subsequent analyses. The second section provides information about the examinees whose scores were used to conduct equating. Background variables are described for the entire sample of examinees available for each exam. Also, the procedures for simulating group differences using background variables are provided in the second section. In the third section, the matching methods are described. The fourth section provides a list of equating designs and methods used in all equating conditions. In the fifth section, four sets of equatings are described: a comparison of SG and CINEG methods, a population invariance study, equating with unmatched groups, and equating with matched samples. Methods for evaluating equating assumptions for the traditional and IRT methods are described in the sixth section, and in the final section, the data analyses described in the first six sections of the methods chapter are connected to the research question(s) they address.

Exams and Scores

Data from the 2007 forms of Advanced Placement (AP) Chemistry, AP French Language, and AP Physics B, and the 2008 form of AP English Language, were chosen for this dissertation because these exams had relatively large sample sizes and represent several different content areas. AP English Language had a very large sample size which was useful for generating large differences between groups, as discussed later. AP French Language included examinees that were both native speakers and classroom language learners. To be consistent with operational equating procedures, only classroom language learners from the AP French Language data set were used in this dissertation.

Although AP Exams were used in this dissertation, several modifications were made to the scores and examinee groups. Many of these changes might make direct inference from the results of this dissertation to the AP Exams impossible. In fact, the College Board is currently undergoing an AP Program review process that is likely to change the nature of the scores in subsequent forms of these exams. Therefore the focus of this dissertation is on the application of equating methods generally, and not on the specific implications of the finding for AP Exams.

Currently the MC items for AP Exams are operationally scored with a correction for guessing, also called formula scoring. Use of formula scoring discourages examinees from randomly guessing on items they do not know the answers to. Unfortunately, the use of formula scoring results in the measurement of an extraneous risk taking behavior in addition to the construct of interest (Sherriffs and Boomer, 1954; Slakter, 1968; Votaw, 1936; Ziller, 1957). Additionally, formula scoring tends to result in a substantial amount of missing data which is particularly problematic for IRT methods. An alternative scoring procedure is number correct scoring where an examinee’s score is simply the weighted sum of all correct responses to MC items and all score points obtained for FR items. With number correct scoring there is no penalty for guessing as both skipped items and incorrect items receive a score of zero.

Because of the unnecessary complications that formula scoring involves, number correct scoring was approximated by imputing missing data using a two-way procedure described by van Ginkel and van der Ark (2005). Across the four exams, only 10–32% of examinees responded to all items. In fact, nearly 20% of item responses were missing for Chemistry and Physics B (see Table 3.1). To reduce the amount of missing data, examinees that responded to less than 80% of MC items were eliminated prior to the imputation process. Original sample sizes and sample sizes after the examinee elimination step (80% Response) are shown in Table 3.2. Across the four exams, the examinee elimination step resulted in a reduction in sample size of between 10 and 40%. However, the percentage of missing scores dropped from 7–19% to around 5%. The remaining 5% of missing values were imputed.

As can be seen from Figure 3.1, the distribution of scores in the original data, the 80% response data, and the imputed data were very different. First, the original and 80% response frequency distributions involved formula scores which had noninteger score categories. Examinees obtained negative formula scores when the number of wrong items exceeded that expected by chance. Operationally, negative formula scores were set to zero, which caused a large frequency at the score of zero. The original and 80% response frequency distributions differed mostly at the bottom end of the score scale because the elimination of examinees with less than 80% MC response rates eliminated a large proportion of low scorers.

Although the frequencies for the imputed data appeared much greater than the frequencies for the other two distributions, in fact, the simulation of number correct scores eliminated all noninteger score categories. For example, Chemistry had 301 possible formula score categories (0, 0.25, 0.5, 0.75, 1,1.25,…,75), but only 76 possible number correct score categories (0,1,…,75). With number correct scoring, examinees' scores were not reduced by incorrect responses. In addition, the imputation process resulted in a score equal to or greater than the original number correct score. Therefore, the entire frequency distribution was shifted to the right for the imputed data as compared to the original or 80% response distributions.

The bumpiness of the formula scored frequency distributions compared to the number correct distribution was caused by the formula scoring process which made some score categories "easier" (or more probable) to obtain. Note that the English Language distributions were much smoother than the distributions for the other tests. This is because there was more than 10 times the number of examinees in the English Language data set, and therefore much less sampling fluctuation.

The differences between the original, 80% response, and imputed data sets shown in Figure 3.1 were also apparent when comparing the first four moments in Table 3.2. The means increased when examinees with low MC-response rates were excluded and increased again when the data were imputed and formula scoring was not used. The imputed data tended to be less variable and more negatively skewed (or less positively skewed) than the other two data sets.

Note that the MC scores in Table 3.2 were not weighted by the section weights. 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. Many psychometric procedures and programs are designed for integer scores. To avoid rounding noninteger scores, integer weights were used in this dissertation. Integer weights were selected so that the MC and FR contributions to the composite score were similar to the operational contributions. The use of integer weights resulted in different ranges of composite scores than those obtained with noninteger weights. The integer weights used in this dissertation and the noninteger weights used operationally, along with the corresponding score ranges are provided in Table 3.3.

Construction of Forms A and B

In order to test the strong assumptions of the various equating methods used with the CINEG design, it is necessary to have examinee responses to both the old and new forms. However, in practice the CINEG design is used when examinee responses are only available for one of the forms. One way to resolve this issue for research purposes is to split a single form into two subforms with common items. The two subforms could be equated using a SG design because the first subform and the second subform are taken by the same examinees. The subforms can also be equated using CINEG methods because the two subforms share common items. Chained equipercentile and frequency estimation equating assumptions can be directly assessed because both “new form” and “old form” score distributions are observed in “both groups”.

The four AP Exams included in this dissertation were all considered new form exams. Each of the four forms was divided into two subforms which are hereafter referred to as Form A and Form B, to distinguish them from the usual notation used for full-length exam forms (X and Y). 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 (see Figure 3.2). The use of “old” and “new” is not particularly relevant for items taken at the same time, but these terms mirror the usual descriptions and notation used with the CINEG design. Note also that the groups here are equivalent because the same examinees took both forms. Therefore using the CINEG design with these data would not make sense operationally, but this artificial situation makes it possible to test the statistical assumptions used with nonequivalent group equating methods.

Operationally, common item sets contained only MC items. The operational MC common items, linking the new forms used in this dissertation to old forms not included in this dissertation, were selected as common items for Forms A and B. Using the entire operational common item set with the shorter Forms A and B resulted in a larger ratio (0.4–0.5) of common items to total MC items than would typically be seen with AP Exams in practice (0.2–0.4). However, items were not removed from the operational common item set because they should have already been chosen to be as similar as possible to the MC items on the whole exam form in terms of content and statistical properties. Also, inclusion of FR items in the common item set is an important consideration, but it is beyond the scope of this dissertation.

Noncommon MC items were assigned to Forms A and B using two content subcategorizations. For example, in AP Chemistry, one content category was “Thermodynamics”. Within the Thermodynamics category were several subclassifications including “First Law”. Additionally, testlets, or groups of items tied to the same stimulus, were typically kept together unless 1) it was necessary to subdivide to try to keep the number of items within Forms A and B the same, and 2) the content categorizations for items within the testlet were different. Occasionally it was necessary to add the same operational non-common items to both Forms A and B so that they would contain equal numbers of items. However, because the operational common item set was considered optimal, these additional items were treated as noncommon items when performing CINEG equating. Following all of these constraints, MC items were assigned to forms on a random basis. For example, if there were four discrete items (not part of a testlet) within the same content categories, then two of the items were randomly assigned to Form A and two were randomly assigned to Form B using a random number generator. Modifications were made to MC item assignment to ensure that the forms contained equal numbers of items.

FR items were assigned to one form or the other based on their difficulty and weight. For example, AP Chemistry had six FR items with integer weights 3, 3, 3, 1, 2, 2, category ranges of 0-9, 0-10, 0-10, 0-15, 0-9, 0-9, and mean as a percent of max values of 50, 39, 37, 51, 36, and 40. Therefore two items with a weight of 3, the item with a weight of one, and one item with a weight of 2 were assigned to each form, while simultaneously making sure to keep the overall FR difficulty levels the same. With these constraints, items 1, 2, 4, and 5 were assigned to Form A and items 1, 3, 4, and 6 were assigned to Form B. For AP English Language there were only three FR items so they were assigned to both forms, but treated as noncommon items for equating purposes.

Modifications to the assignment of items to forms would have been made to ensure that the forms were of approximately equal difficulty, but the weighted MC, FR, and composite means were fairly close for all four exams, as can be seen from Table 3.4. The composite score is the sum of the integer weighted MC and FR sections (for section weights see Table 3.3). The numbers of MC, common, and FR items in the original AP forms, and assigned to Forms A and B are provided in Table 3.5. Item numbers for the Form A noncommon items, Form B noncommon items, and common items are provided in Table 3.6. Bold item numbers indicate that the noncommon item was assigned to both Form A and Form B but was not included in the common item set for equating purposes. English Language has an especially large number of noncommon items that were assigned to both forms because the items were part of a large testlet. Because there was no obvious way to split the items and maintain content balance, the items were kept together.

Samples

Background Variables in Original Data

Background information about the examinees was available for gender, ethnicity, parental education, examinee grade level (e.g., 10th grade), a fee reduction indicator for low-income examinees, and region of the country. Ethnic groups of notable size included (1) African American, (2) Mexican or Mexican American, (3) Puerto Rican, Latino, or other Hispanic, (4) Asian, and (5) White (missing values were coded 0). The four regions of the country: (1) Northeast, (2) South, (3) Midwest, and (4) West, were coded based on the state provided for the examinee and the US census definition of the regions (*www.census.gov*). Parent education was defined as the highest level of education obtained by either parent. Levels were coded into five categories:

0. Missing
1. Any education through high school diploma or equivalent
2. Trade school, some college, or associates degree
3. Bachelors degree or some graduate/professional school

4. Graduate/Professional degree

Three grade levels were considered in this dissertation because the majority of examinees fell into (1) 10th, (2) 11th, or (3) 12th grade (missing values were coded 0). Fee indicator was a dichotomous variable equal to 1 for low-income students that received a reduction in the cost of the exam and 0 for anyone else. For gender, males were coded 1, and females were coded 2. Apart from region and fee indicator, all variables were self-reported and therefore included a certain proportion of missing values and an unknown proportion of incorrect values. However, 100% of examinees provided their gender, but a small percentage of examinees had no reported region for AP English Language.

The distribution of background variables in Table 3.7 indicates that a higher proportion of males took AP Chemistry and AP Physics B, but substantially more females took the AP French Language and AP English Language Exams. Also, the percent of examinees that received an exam fee reduction was 7–10% across all exams. The percentage of examinees within each of the four regions varied somewhat across exams;Region 2 (the South) took the most AP Exams and Region 3 (the Midwest) took the fewest AP Exams in general. The percentages of examinees within each grade level indicated that most examinees took these four AP exams in grades 11 or 12. Very few examinees did not respond to the parental education background question (2–5% nonresponse). As the level of parental education increased, the corresponding percentage of examinees also tended to increase. At least 30% of AP examinees had one or more parents with a graduate level degree. About 3% of examinees failed to identify their ethnicity. White examinees were the majority examinee group. Asian examinees comprised 11–22% of examinees, and all of the other ethnic groups had comparatively low percentages of examinees.

Simulating Group Differences

A main objective of this dissertation is to investigate the impact of group differences on equating assumptions and properties. However, the group performance differences across operational AP forms are quite small, and the distribution of background variables is fairly consistent across administrations. The effect size (ES) across years for scores on the common items is not shown here but tends to be less than 0.1 and is often near zero for many of the AP Exams. ES was calculated:

$$ES = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}}, \qquad (15)$$

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. Forms A and B were taken by the same examinees. Therefore the common item ES was exactly zero, because the common item scores were the same for the two forms. According to Kolen and Brennan (2004):

> …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. (p. 286)

With small group differences, and old and new form groups that are very similar in terms of background variables, there is little need to consider matched samples equating—the groups are naturally matched. With Forms A and B, there were no group differences—the groups were identical (matched exactly). However, the College Board has considered moving to multiple annual administrations. Groups tested at different times of the year often differ in performance (e.g., Schmitt, Cook, Dorans, & Eignor, 1990). Matched samples equating would become a relevant consideration for the AP Exams should the College Board adopt additional administration dates.

In order to investigate the research questions and use the available data sets, pseudo old and new form groups were sampled from the original group that took Forms A and B using a background variable related to common item scores. Out of the six background variables considered in this dissertation, 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.

The impact of sampling from parental education categories to simulate group differences can be predicted by comparing the common item means across the four education categories. Table 3.8 provides the means and standard deviations for the integer weighted common item section for each of the parental education categories, excluding the nonresponse category. Note that the common items were weighted by the MC section weights provided in Table 3.3. The sample sizes in each category ranged from nearly 500 to nearly 3500 for the exams with smaller sample sizes and from 10,000 to over 40,000 for AP English Language. In general, performance on the common item section increased as parental education increased. A comparison of the mean differences between the highest and lowest categories indicated that the group means differed by more than a standard deviation in some cases. The nonresponse sample sizes (not shown here) were fairly small in comparison and the means were slightly below the weighted average of the common item means for the other four categories across all exams.

In order to study the impact of group differences on equating results, assumptions, and properties, four common item ESs were considered. Four separate pseudo old and new form groups were constructed by sampling examinees from the four levels of parental education, in order to achieve the desired ESs. For the three exams with smaller sample sizes (AP Chemistry, AP French Language, and AP Physics B), it was not possible to create nonequivalent pseudo groups, match the pseudo groups, and obtain a sample size of at least 1500 per form, for ESs larger than 0.3. Therefore, for these three exams, the four ESs considered were 0, 0.1, 0.2, and 0.3. However, for AP English Language, which had a much larger original sample, more extreme ESs were considered including 0, 0.25, 0.5, and 0.75.

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. For example, out of 12,328 AP Chemistry examinees, 11,926 had nonmissing parental education values. A random sample of 525 examinees in the first parental education category was assigned to Form A; the other 525 examinees were assigned to Form B. The same procedure was used for examinees in the other three parental education categories (see Table 3.8). For parental education categories with an odd number of examinees, the extra examinee was assigned arbitrarily to Form B. Stratified random sampling, using the SAS SurveySelect Procedure, automatically resulted in “old” and “new” form pseudo groups with common item ESs near zero.

To obtain ESs greater than zero, the ratios of examinees within parental education categories were manipulated using two iterative steps. First, the number of examinees within each level of parental education was adjusted based on the means presented in Table 3.8. Sample-size weighted means were computed after each adjustment to determine the effect of the adjustments on the common item ES. Adjustments to the numbers of examinees for Form A and Form B were kept the same. In other words, if 1200 examinees were removed from some levels of parental education for Form A, then 1200 examinees were also removed from some levels of parental education for Form B. Sample sizes were also kept as large as possible across the four levels to ensure that matched samples would have at least 1500 examinees per form. Adjustments continued until the computed ES was within 0.01 of the ES of interest. When an acceptable ES was obtained, the numbers of examinees within each category were entered into the SAS SurveySelect Procedure. The second iterative process involved running the SurveySelect

Procedure for Forms A and B. After the two samples were generated, the ES was calculated. The process was rerun, modifying as necessary the number of examinees sampled from each level of parental education, until the resulting ES was again within 0.01 of the desired ES. When the desired ES was obtained, the process was complete with the creation of data sets for the pseudo groups for both forms.

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 (see Figure 3.3). Because of the relationship between examinee performance and parental education, this sampling technique resulted in higher performing pseudo old form groups.

Using data from AP Chemistry as an example, to obtain a desired ES of 0.3, examinees in the higher parental education levels were removed from Form A, and examinees in the lower parental education levels were removed from Form B, until the computed ES was near 0.3. After several iterations, the numbers of examinees that produced the desired ES were obtained, and are presented in Table 3.9 (see Form A Pseudo and Form B Pseudo). The sample sizes were reduced by 537 for parental education level 3 and by 1688 for parental education level 4 for Form A. For Form B, 375 examinees were removed from parental education level 1, 624 were removed from parental education level 2, 1226 examinees were removed from parental education level 3, and two were removed from parental education level 4 (to ensure equal pseudo sample sizes for both forms). After a few iterations of SAS SurveySelect, an ES of 0.30255 was obtained which was very close to the desired ES of 0.3.

The effect of the sampling process on sample size can be seen by comparing the total sample sizes. The original Forms A and B had approximately 5962 examinees, and the pseudo groups with an ES of 0.3 each had 3737 examinees. Matched sampling, which is discussed in the next section, also resulted in a reduction in sample size. In order to keep the equating results comparable in terms of sampling error, a sample of 1500 was taken from each of the pseudo and matched samples for all exams. Figure 3.4 provides a graphical example of the samples for AP Chemistry. The largest pseudo samples were obtained for the zero ES, and the sample size decreased as the ES increased. The smallest samples were the matched samples where matching was applied to pseudo groups with ESs larger than zero. As indicated in Table 3.9, the Chemistry sample size for an ES of zero was 5962, the sample size for an ES of 0.3 was 3737, and the matched samples where matching was performed on the pseudo group with an ES of 0.3 was 2050. The black ring within each pie chart in Figure 3.4 represents the subsample of 1500 that was taken out of each of the pseudo and matched groups for equating.

Desired and observed ESs and the percentage of examinees within each parental education category are listed in Table 3.10 for all four exams. Examinees with missing values for parental education were not included in the pseudo groups. The observed ESs for the original pseudo samples (Obs ES_1) were very close to the desired ESs. The observed ESs based on the samples of 1500 (Obs ES_2) were still close to the desired ESs, but not as close as the original observed ESs.

Matching Methods

Matching old and new form groups has typically involved background variables, examinee scores on other related measures, or common item scores (e.g., Cook, Eignor, & Schmitt, 1988; Lawrence & Dorans, 1990; Schmitt, Cook, Dorans, & Eignor; Way, Davis, Fizpatrick, 2006; etc.). Operationally, the factors that differentiate group membership, called selection variables, are unknown. However, in this dissertation, to simulate group differences, the pseudo groups were created based on a known selection variable (i.e., parent education level).

Matching the pseudo groups on the selection variable represents the best case scenario for operational matching, and was expected to produce equating results similar to those produced with the original new and old form groups that were equal in performance. In fact, matching the pseudo groups on the selection variable is similar to taking a random sample from the original old and new form groups. The sample size of the matched data (or the random sample) may be much smaller than the original sample size because some examinees were excluded to create pseudo groups, and more examinees were excluded when matching.

Equating results and assumptions were studied under the following four matching conditions:

1. No Matching (M_0): Here equating was conducted using the pseudo groups created by sampling from parental education categories to achieve the desired ES. No matching was done.
2. Matching on Selection Variable (M_1): Pseudo groups were matched by randomly selecting the same number of examinees within each category of parental education. For example, if the sample sizes in the first parental education category were 1000 and 2000 for pseudo groups A and B respectively, then the largest number of examinees that could have been matched was 1000. Therefore, all 1000 examinees were selected from pseudo group A and 1000 examinees were randomly selected out of the 2000 available for pseudo group B. The same process was repeated for parental education categories two through four.
3. Matching on Propensity Score Including Selection Variable (M_2): Pseudo groups are matched on a propensity score obtained from a logistic regression equation including all six background variables (see Table 3.7). Matching was done using a SAS macro developed by Parsons (2001). The SAS Logistic Procedure was used to calculate propensity scores for each examinee. Those

examinees in Form A whose propensity scores matched the propensity scores of Form B examinees based on five decimal places of precision were matched first. The program then matched Form A and Form B examinees whose propensity scores matched based on four decimal places of precision. The process continued until all Form A and Form B examinees were matched that had propensity scores within one decimal place of precision. Examinees that did not have a match based on this process were not included in further analyses.

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

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

Equating Designs and Methods

The two equating designs considered in this dissertation 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. Smoothing parameters were chosen based on a visual inspection of the smoothed and unsmoothed equivalents. The smoothing value that provided the smoothest equating relationship within a band of one SE around the unsmoothed equivalents was chosen. For the IRT observed score and frequency estimation methods, synthetic weights of 0.5 for both new and old form groups were used. Equations and assumptions for all equating designs and methods included in this dissertation are provided in Chapter 2 (see also Kolen & Brennan, 2004).

MULTILOG (Thissen, 1991) was used for IRT estimation for the 3PL model and the GRM. For IRT observed score equating, normal quadrature points 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 but not for the CINEG design. The *Equating Recipes* scale transformation code for the Haebara method (1980) was used for the CINEG design to place parameter estimates on the same scale before equating.

SEs were estimated using the bootstrap procedure with 1000 replications. Using C functions in *Equating Recipes,* equating results and SEs for traditional equating methods were calculated using the same smoothing value for each replication. IRT SEs were not estimated because they are not estimated by *Equating Recipes*, and using the bootstrapping process with other programs is too time consuming to be practical for the number of conditions involved in this dissertation.

Comparison of Equating Results

Four sets of equating analyses are described in the next four sections. In the first set of analyses, results were compared for equating methods using the same data but two different designs (SG and CINEG). The second section provides a description of a population invariance study where levels of the selection variable were used as subgroups. The third section describes equating methods and comparisons used with unmatched nonequivalent groups. The final section describes equating methods and comparisons used with different matched samples equating procedures.

SG and CINEG Equating Methods

Because Forms A and B were created by splitting one form into two, examinee responses to both forms were available. Two forms taken by the same examinees could be equated using a SG equating method like the direct equipercentile or IRT true or observed score equating. However, since Forms A and B were also constructed to share common items, equating could be done using the CINEG equating methods like frequency estimation, chained equipercentile, or IRT true or observed score. As mentioned in Chapter 2, von Davier, Holland, and Thayer (2003, 2004) proved that the frequency estimation and chained equipercentile methods would produce the same results when the common item score distributions are the same in both groups. This extreme case holds for the SG design and therefore for Forms A and B.

As a baseline comparison, Forms A and B were equated using three SG nonlinear equating methods: direct equipercentile, IRT true score, and IRT observed score. Then the forms were equated using four nonlinear equating methods used with the CINEG design: frequency estimation, chained equipercentile, IRT true score, and IRT observed score. The results obtained from all seven methods were expected to be nearly identical because the groups were equivalent. The resulting seven equating relationships were compared for all four AP Exams. The only criterion for these comparisons was the consistency of results.

Population Invariance

The equating property of population invariance holds when the equating relationship is the same for all examinee groups. Equating results are considered population dependent when different equating relationships are obtained for different groups. 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). If the equating relationships do not differ by some specified threshold (e.g., the DTM), then the population invariance property is considered to have held for the given data and groups considered. In this type of population invariance study, groups may differ substantially from one another in terms of total score ESs. 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. When population invariance is assessed using the SG design, the female group is identical for both forms. Therefore, the equatings conducted in a population invariance study involve very similar (or the same) 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 four AP Exams, equating relationships were estimated for each level of parental education using the SG design. As in the typical population invariance study, equating was conducted separately for each level of parental education. For example, Form A was equated to Form B using only examinees in parental education level four. The equating relationship between the two forms was similarly estimated for other levels, and then the resulting equating relationships were compared. For AP Chemistry, French Language, and Physics B, a sample of 1500 was taken from each of the levels. Because the first two levels were smaller than 1500, they were collapsed into a single category, then a sample of 1500 was taken for equating. As a “population” or total group equating, a stratified random sample of 1500 was taken from across all parental education levels. Therefore, for the three exams with smaller sample sizes, four equating relationships were compared: one based on the combined levels 1 and 2, one based on level 3, one based on the fourth level, and finally, one based on a stratified random sample across all levels. For AP English, separate equating relationships were estimated for all four parental education categories and for a stratified random sample across all levels of parental education (the total group equating relationship).

Because the purpose of this analysis was to determine the extent to which equating results were population dependent for the four exams, only the SG design was used so as not to confound population invariance with violations of 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 samples so that it was possible to compare the equating relationships based on the parental education levels to the total group, taking sampling error into account. 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. Two SEs were used following the same logic as usual significance testing. Based on the normal distribution, the population parameter (in this case the true equating relationship) should fall within two standard errors of the sample value 95% of the time. The SEs for the direct equipercentile equating method do not apply strictly to the IRT equating results, but they serve as a useful baseline. Some research indicates that IRT SEs tend to be smaller than traditional SEs (Liu & Kolen, 2010), which would mean comparing two IRT equating relationships based on two traditional SEs would be anti-conservative. Additionally, the standard error of the equating difference (SEED; von

Davier, Holland, & Thayer, 2004b) would have been a more appropriate statistic for comparing the difference between subgroup and total group equating relationships because it takes into account sampling error in both equating relationships. Because the SEED takes into account more sources of variability than the SE, SEs are likely to be smaller than SEED values would have been. Therefore SEs may be too conservative for determining whether equating relationships are within sampling error of one another. 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.

REMSD and classification consistency of AP grades (1 to 5) were also calculated to compare total group and subgroup equating relationships. Classification consistency 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.

$$CC = \frac{n_{11} + n_{22} + n_{33} + n_{44} + n_{55}}{N}, \tag{16}$$

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 0.5 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 French 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 composite cut score for AP French Language Form B.

This procedure resulted in approximately equal numbers of examinees receiving each AP grade under formula scoring and under simulated number correct scoring. A change to number correct scoring operationally would require the College Board to conduct a new grade setting and adopt new cut scores for these exams. For political reasons, it seems unlikely that cut scores would be adopted that would result in extremely different proportions of examinees receiving each AP grade. Therefore the cut score placements used here appear reasonable even though results based on these cut scores may not be directly generalized to the current AP Exams.

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

With three SG design equating methods and three exams with three equating relationships based on parental education levels and one equating relationship based on the total group, there were 36 equating relationships, not including those for AP English Language. For English Language there were three equating methods, four equating relationships based on parental education levels, and one equating relationship based on the total group, for a total of 15 equating relationships. The equating relationship based on the total group was considered the criterion equating relationship for these analyses. Each comparison equating relationship was compared to the criterion equating relationship obtained using the same equating method.

Equating with Different Effect Sizes (M_0)

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.

To compare equating results with groups that vary in exam performance, old and new form pseudo groups were created by changing the ratio of examinees within each level of parental education in the two forms. Pseudo groups were created to systematically differ by an ES of approximately 0, 0.1, 0.2, and 0.3 for exams with smaller sample sizes, or 0, 0.25, 0.5 and 0.75 for AP English Language. Equating is known to work best when groups perform similarly (Kolen, 1990). Therefore, equating results based on pseudo 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. Note that SEs were not calculated for the IRT methods because of the computation time required. The SEs for the traditional methods were used as a reference but are not strictly applicable to the IRT equating results. As in the population invariance comparisons, the SE was used for comparisons rather than the SEED because the SEED is not part of the standard functions in *Equating Recipes*.

Although a graphical comparison is helpful for understanding if equating relationships change based on group differences, REMSD was used to provide an overall indication of how equating relationships departed from the criterion relationship as ES increased.

$$REMSD = \frac{\sqrt{\sum_{\min(a)}^{\max(a)} v_{ac}[eq_{Bc}(a) - eq_{B0}(a)]^2}}{\sigma_0(B)}. \qquad (17)$$

Here v_{ac} is the conditional proportion of Form A examinees at a particular Form A score for the comparison (c) equating relationship; the Form B equivalent for the pseudo group with ES of zero is $eq_{B0}(a)$; and the Form B equivalent for a comparison pseudo group with an ES greater than zero is $eq_{Bc}(a)$. The summation is taken over all Form A scores. The equating method used with the criterion pseudo group (ES = 0) was also used with the comparison pseudo group (ES = 0.1, 0.2, 0.3 or 0.25, 0.5, 0.75).

As in the population invariance studies, the practical significance of using an equating relationship with groups that differ substantially in performance was quantified using classification consistency of AP grades. Classification consistency was calculated using the grades new group examinees received given the criterion equating relationship (ES = 0), and the AP grades they received with a comparison equating relationship (ES > 0).

With four ES levels, four exams, and four equating methods, there were 64 equating relationships estimated. Again, each comparison equating relationship was compared to the criterion equating relationship using the same equating method. The SG design equating results for each exam, described previously, might also have been considered criteria because the ES difference between groups was zero. In fact, if population invariance holds, and the equating assumptions of the CINEG equating methods hold, then the equating relationship found with the pseudo group with an ES of zero should be very similar to the SG design equating results.

Equating with Matched Samples (M_1–M_3)

Based on findings from previous research (Harris & Kolen, 1990; Sinharay & Holland, 2007) it is expected that as the ES increases, the equating results will become farther away from the criterion equating relationship. The last set of analyses were based on unmatched data (M_0) with ESs ranging from zero to 0.75. The current set of analyses used three different matching techniques (M_1–M_3) to make dissimilar groups more similar. The efficacy of matching in improving equating accuracy was determined by comparing the matched samples equating results to the criterion equating relationship with an ES equal to zero. Matched samples equating was conducted using four equating methods: frequency estimation, chained equipercentile, IRT true score, and IRT observed score. As in the M_0 comparisons, each matched samples equating relationship was compared to the criterion equating relationship based on the same equating method. The SE calculated for the criterion equating relationships in the M_0 analyses was also used to interpret differences in the criterion and matched samples equating results. REMSD and classification consistency were used to quantify the equating accuracy of the matched samples equating results. The improvement in equating accuracy with matched sampling over unmatched equating was investigated by comparing the REMSD and classification consistency indices for M_0 versus M_1–M_3.

Matching is not necessary when there are no group differences. Therefore the three matching methods were applied only to the three sets of pseudo groups with ESs greater than zero. With three ESs, three matching methods, four equating methods, and four exams, 144 equating relationships were estimated.

Evaluation of Equating Assumptions

Two ways to assess assumption violations are 1) direct evaluation of the statistical assumptions and 2) comparisons of equating results for different equating methods.

Direct Evaluation of Statistical Assumptions

Direct evaluation of the statistical assumptions of various equating methods can be done by calculating typically unobservable values like the conditional distribution of the total score on Form A, given the common item score, for both the old and new form groups. Direct evaluation of traditional equating methods used with the CINEG design can only be done when the responses of examinees have been observed for both Forms A and B. Examinee responses were observed for both forms in this dissertation because a single operational form was divided into two “half-tests”.

Frequency Estimation Method

The assumptions for the frequency estimation equipercentile method are as follows:

1. The conditional distribution of Form A composite scores given common item scores is the same in both old and new form groups.
2. The conditional distribution of Form B composite scores given common item scores is the same in both old and new form groups.

Direct evaluation of frequency estimation method assumptions involves comparisons of v +1 pairs of conditional distributions, where v is the number of common items. In order to quantify how well the frequency estimation assumptions are met, 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. Each of the $v + 1$ maximum differences in the cumulative frequency distributions was similar to a Kolmogorov-Smirnov statistic. However, differences in the cumulative frequencies were indexed in terms of cumulative frequencies rather than cumulative proportions.

Chained Equipercentile Method

The assumptions for the chained equipercentile method are as follows:

1. The linking relationship from Form A composite scores to the common item scores ($e_V(A)$) is the same for both old and new form groups.
2. The linking relationship from the common item scores to Form B composite scores ($e_B(V)$) is the same for both old and new form groups.

For the evaluation of chained equipercentile method assumptions, REMSD was calculated for the equating relationships ($e_V(A)$) and ($e_B(V)$) in the old (2) and new (1) form groups:

$$REMSD\big(e_V(A)\big) = \frac{\sqrt{\sum_{\min(a)}^{\max(a)} v_{ac}[eq_{V1}(a) - eq_{V2}(a)]^2}}{\sigma(V)_2},$$

$$REMSD\big(e_B(V)\big) = \frac{\sqrt{\sum_{\min(v)}^{\max(v)} v_{vc}[eq_{B1}(v) - eq_{B2}(v)]^2}}{\sigma(B)_2}. \qquad (18)$$

Unsmoothed equipercentile equivalents were calculated for the link from *A* to *V* and from *V* to *B* for both old and new form groups to assess the chained equipercentile method assumptions.

IRT True and Observed Score Methods

IRT true and observed score equating methods also have statistical assumptions which can be evaluated, albeit less directly. The following unidimensional IRT equating assumptions were evaluated:

1. The construct measured by the exam is unidimensional. This assumption and the next (2) were tested using a variety of correlational analyses as described below.
2. The same construct is measured by both Form A and Form B and in both the old and new form groups.
3. 3PL and GRM fit the data. Examination of item parameter estimates for evidence of model fit.

To investigate whether or not IRT model assumptions held for the data in this dissertation, MC-FR correlations (uncorrected and disattenuated using coefficient alpha reliability estimates), principal component analysis on polychoric correlations, and results of PolyDIMTEST (Li & Stout, 1995) and PolyDETECT (Zhang, 2007) were considered. Low MC-FR correlations (especially disattenuated correlations) and scree plots that indicated more than one important factor may indicate that unidimensional IRT assumptions did not hold. Significant test statistics for PolyDIMTEST and multiple item clusters for PolyDETECT also indicate lack of unidimensionality. Likewise, IRT estimation iterations that failed to converge, item parameter estimates that appeared out of bounds, or item parameter estimates that changed greatly when estimated with and without other item types, suggest that the IRT calibration and equating process may be inaccurate. 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. The degree to which the IRT assumptions held or failed to hold based on all of the above criteria was used to inform the interpretation of IRT equating results.

Comparison of Equating Results

Equating results obtained using CINEG equating methods were compared to those obtained using the SG equating methods. Equating methods developed for use with the SG design have minimal assumptions. Therefore substantially different results for CINEG and SG methods would be caused presumably by a violation of CINEG method assumptions.

Addressing the Research Questions

There are five sets of analyses included in this dissertation:

1. A comparison of equating results using SG and CINEG design equating methods
2. A population invariance study using levels of parental education to define different examinee subgroups
3. A comparison of equating results using a variety of ESs for old and new form groups (M_0)
4. A comparison of equating results using three different matched samples equating techniques (M_1–M_3)

5. An evaluation of traditional and IRT equating method assumptions

The first set of analyses was used as a baseline check, to make sure that the equating results obtained using SG and CINEG equating methods provided reasonable and similar results. Analysis sets two through five were used to answer one or more of the research questions. The first research question, *are equating results invariant for populations of examinees with different levels of parental education*, was addressed by the second set of analyses. The second research question, *what is the impact of group differences on equating results*, was addressed by the unmatched sampling condition (M_0) in the third set of analyses. The third research question, *what is the impact of group differences on the degree to which equating assumptions are met*, was addressed by a combination of the results of the third and fifth sets of analyses. The fourth research question, *which matching techniques, if any, provide more accurate equating results*, was addressed by a combination of the third and fourth sets of analyses. Finally, the fifth research question, *can matched samples equating reduce the extent to which equating assumptions are violated*, was addressed by a combination of the third, fourth, and fifth sets of equating analyses.

Analyses 1 through 4 involved 287 equating relationships. Table 3.11 provides a list of the equating relationships for each analysis set along with the criteria used to evaluate the equating results.

Summary

This chapter provided a summary of the methods that were employed to answer the five research questions of interest. The exams chosen included the 2007 forms of AP Chemistry, AP French Language, and AP Physics B, and the 2008 form of AP English Language. The rationale for exam selection was given as well as the process for creating subforms, adjusting scoring procedures, selecting common items sets, and creating pseudo groups.

Parental education was selected from six potential background variables to create pseudo groups that differed in exam performance; background variables were also used to adjust for differences in performance in the pseudo groups using a variety of matching techniques. The two equating designs used in this dissertation, the SG design and the CINEG design were described as well as the various traditional and IRT methods that were used to conduct equating. *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009) was the primary software used to conduct equating for this dissertation.

Two methods for assessing the violations of equating assumptions were described in this chapter: direct comparison and equating methods comparison. Also considered were a variety of criteria for assessing population invariance including REMSD statistics, SEs, and classification consistency.

Finally, this chapter provided a list of procedures that were used to compare the results of, population invariance studies, equating with group differences, and matched samples equating. These research questions were designed to provide a comprehensive look at how equating assumptions and properties are affected by group differences, and how successful matching procedures are at mitigating the effects of group differences.

Table 3.1 Percentage of Examinees with Missing MC Data

Exam	% of examinees with no missing responses	% missing values before examinee deletions	% of missing values after examinee deletions
Chemistry	12.21	17.5	5.7
French Language	31.82	6.9	4.3
Physics B	10.59	18.8	5.4
English Language	30.92	10.5	4.4

Table 3.2 Unweighted MC Moments in Original, 80% Response, and Imputed Data

Exam	Data	N	M	SD	Skew	Kurt
Chemistry	Original	20,000	28.74	16.41	0.32	-0.66
	80% Response	12,328	33.69	17.22	-0.07	-0.83
	Imputed	12,328	44.54	15.40	-0.30	-0.94
French Language	Original	15,212	37.13	19.25	0.06	-0.81
	80% Response	13,907	38.56	19.17	-0.02	-0.79
	Imputed	13,907	51.51	15.07	-0.18	-0.77
Physics B	Original	20,000	23.38	14.51	0.45	-0.51
	80% Response	12,577	27.05	15.46	0.14	-0.82
	Imputed	12,577	37.12	13.98	-0.00	-1.04
English Language	Original	301,095	29.37	12.05	-0.25	-0.58
	80% Response	247,197	31.77	11.36	-0.45	-0.21
	Imputed	247,197	38.12	9.48	-0.74	0.12

Table 3.3 Integer and Non-Integer Weights and Scales

Exam	Section	N Items (Categories)	Integer Weights	Non-Integer Weights
Chemistry	MC	75 (0-1)	2	1
	FR	1 (0-9)	3	1.6666
	FR	2 (0-10)	3	1.5
	FR	1 (0-15)	1	0.5
	FR	2 (0-9)	2	1.25
	Scale		0-288	0-150
French	MC	42 (0-1)	2	0.9523
	MC	43 (0-1)	2	0.9302
	FR	30 (0-1)	1	0.5333
	FR	1 (0-9)	5	2.6666
	FR	5 (0-5)	3	1.6
	Scale		0-320	0-160
Physics B	MC	69 (0-1)	1	1.3043
	FR	2 (0-15)	1	1.125
	FR	5 (0-10)	1	1.125
	Scale		0-149	0-180
English	MC	55 (0-1)	2	1.2272
	FR	3 (0-9)	5	3.0556
	Scale		0-239	0-150

Table 3.4 Weighted Means and SD for MC, CI, FR, and Composite Scores (After Imputation)

			MC		CI		FR		COMP	
Exam	Form	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Chemistry	A	12,328	59.07	20.03	31.60	11.08	44.45	23.21	103.51	41.53
	B		61.62	21.95			44.53	24.22	106.14	44.47
French	A	13,907	69.17	20.93	31.74	10.19	61.22	18.28	130.39	37.00
	B		67.42	19.68			60.75	18.70	128.17	36.14
Physics B	A	12,577	24.89	9.36	11.95	4.79	18.36	10.50	43.25	18.94
	B		24.72	9.63			17.85	9.99	42.57	18.56
English	A	247,197	55.92	13.95	21.20	6.60	69.97	19.59	125.88	29.89
	B		54.39	14.71					124.35	30.52

Note. Score means are not standardized so differences in their magnitudes across exams do not indicate the relative difficulty of the exams.

Table 3.5 Exam and Group Information

		MC Items		FR Items
Exam	Form	Total	Common	Total
Chemistry	2007	75	25	6
	A	50	25	4
	B	50	25	4
French Language	2007	85	26	36
	A	56	26	19
	B	56	26	19
Physics B	2007	69	22	7
	A	46	22	5
	B	46	22	5
English Language	2008	55	16	3
	A	40	16	3
	B	40	16	3

Table 3.6 Structure of Forms A and B Noncommon and Common Items

Exam	Form A Non-Common Items	Common Items	Form B Noncommon Items
Chemistry	8,9,10,11,12,16,26,28,29,31,39,41,45,50, 56,57,58,59,64,66,68,69,71,73,75,**FR1**, FR3,**FR4**,FR6 (25MC 4FR)	4,5,6,7,13,14,15,19,21,23, 27,32,33,34,35,38,44,46,49, 51,53,60,65,67,70 (25CI)	1,2,3,17,18,20,22,24,25,30,36,37,40,42, 43,47,48,52,54,55,61,62,63,72,74,**FR1**, FR2,**FR4**,FR5 (25MC 4FR)
French	**1**,3,4,9,16,18,19,21,22,23,24,25,26,43,44, 45,46,47,48,49,50,51,52,79,80,81,82,83, 84,85,FR1-FR15**,FR31,FR32**,FR33,FR35 (30MC 19FR)	5,7,13,14,20,37,38,39,40,41, 42,53,55,56,58,59,60,62,63, 64,65,66,67,68,69,70 (26CI)	**1**,2,6,8,10,11,12,15,17,27,28,29,30,31, 32,33,34,35,36,54,57,61,71,72,73,74,75, 76,77,78,FR16,FR30,**FR31**,**FR32**,FR34, FR36 (30MC 19FR)
Physics B	1,6,11,12,14,15,26,29,33,34,35,**37**,40,42, 43,44,47,49,50,51,52,56,58,63,**FR1**,FR3, **FR4, FR5**,FR7 (24MC 5FR)	4,5,9,10,13,16,17,21,25,28, 31,32,36,38,41,45,46,61,64, 66,67,68 (22CI)	2,3,7,8,18,19,20,22,23,24,27,30,**37**,39, 48,53,54,55,57,59,60,62,65,69,**FR1**, FR2,**FR4,FR5**,FR6 (24MC 5FR)
English	32,33,34,35,36,37,38,39,40,41,42,43,44, 45,46,**47,48,49,50,51,52,53,54,55,FR1, FR2,FR3** (24MC 3FR)	1,2,3,4,5,6,7,8,9,11,12,13, 14,15,16,17 (16CI)	10,18,19,20,21,22,23,24,25,26,27,28,29, 30,31,**47,48,49,50,51,52,53,54,55,FR1, FR2,FR3** (24MC, 3FR)

Note. Bold items are noncommon items that appear in both Forms A and B.

Table 3.7 Percentage of Examinees in each Background Variable Category

Background Variable	Code	Exam			
		Chemistry	French Language	Physics B	English Language[a]
Gender	1	57	28	68	37
	2	43	72	32	63
Fee	0	92	93	92	90
	1	8	7	8	10
Region	1	23	31	25	14
	2	30	29	30	44
	3	23	16	17	15
	4	24	24	28	26
Grade	0	3	4	4	3
	1	6	3	2	1
	2	54	26	38	82
	3	37	67	56	14
Parent ED	0	3	3	2	5
	1	9	7	9	11
	2	15	11	15	19
	3	33	31	34	35
	4	40	48	40	30
Ethnicity[b]	0	3	3	3	2
	1	4	3	3	6
	2	2	4	3	5
	3	3	5	4	6
	4	22	13	19	11
	5	63	67	64	65

[a] Region percentages do not sum to 100 because of missing values.

[b] These percentages do not sum to 100 because ethnic groups with very small proportions were not included.

Table 3.8 Weighted Common Item Means and Standard Deviations for Categories of Parental Education

		Form A			Form B		
Exam	Category	N	Mean	SD	N	Mean	SD
Chemistry	1	525	23.40	10.80	525	24.38	11.22
	2	924	26.72	11.50	924	27.08	11.26
	3	2025	31.70	10.16	2026	31.58	10.78
	4	2488	35.20	9.78	2489	35.22	9.80
French Language	1	487	25.97	9.90	487	25.89	10.57
	2	750	27.06	9.84	751	27.71	10.04
	3	2148	30.89	9.64	2149	30.76	9.98
	4	3387	34.25	9.61	3387	34.24	9.65
Physics B	1	556	9.03	4.66	557	9.34	4.50
	2	919	10.40	4.56	920	10.26	4.55
	3	2145	11.96	4.58	2145	11.98	4.64
	4	2510	13.18	4.61	2511	13.19	4.59
English Language	1	13,835	17.70	6.54	13,836	17.68	6.58
	2	23,765	19.38	6.50	23,765	19.38	6.48
	3	42,881	21.68	6.24	42,881	21.68	6.26
	4	37,604	23.38	6.08	37,605	23.40	6.04

Note. The common item means are not standardized so differences in their magnitudes across exams do not indicate the relative difficulty of the exams.

Table 3.9 AP Chemistry Sample Sizes of Original, Pseudo, and Rematched Groups

Parental ED	Form A		Form B		Rematched[a]
Level	Original	Pseudo	Original	Pseudo	Forms A & B
1	525	525	525	150	150
2	924	924	924	300	300
3	2025	1488	2026	800	800
4	2488	800	2489	2487	800
Total	5962	3737	5964	3737	2050

[a]Desired ES= 0.3, Obtained ES = 0.30255

Table 3.10 Distribution of Parental Education in Pseudo Groups of Varying ESs

Exam	Group	Desired ES	Obs ES$_1$	Total N	% Cat1	% Cat2	% Cat3	% Cat4	Obs ES$_2$
Chemistry	New	0	0.00939	5962	0.09	0.15	0.34	0.42	0.02033
	Old			5964	0.09	0.15	0.34	0.42	
	New	0.1	0.09777	4974	0.10	0.19	0.41	0.30	0.07507
	Old			4974	0.09	0.15	0.26	0.50	
	New	0.2	0.20713	4074	0.13	0.23	0.40	0.24	0.21124
	Old			4074	0.10	0.10	0.19	0.61	
	New	0.3	0.30255	3737	0.14	0.25	0.40	0.21	0.29137
	Old			3737	0.04	0.08	0.21	0.67	
French Language	New	0	0.00181	6772	0.07	0.11	0.32	0.50	0.01166
	Old			6774	0.07	0.11	0.32	0.50	
	New	0.1	0.09932	5542	0.09	0.13	0.39	0.39	0.11761
	Old			5542	0.05	0.11	0.23	0.61	
	New	0.2	0.20620	4642	0.11	0.16	0.46	0.27	0.20716
	Old			4642	0.08	0.08	0.12	0.72	
	New	0.3	0.29274	4042	0.12	0.19	0.53	0.16	0.31743
	Old			4042	0.06	0.06	0.12	0.76	
Physics B	New	0	0.00417	6130	0.09	0.15	0.35	0.41	0.04002
	Old			6133	0.09	0.15	0.35	0.41	
	New	0.1	0.09697	5024	0.11	0.18	0.41	0.30	0.08471
	Old			5024	0.08	0.11	0.31	0.50	
	New	0.2	0.20186	4024	0.14	0.23	0.43	0.20	0.22676
	Old			4024	0.08	0.10	0.20	0.62	
	New	0.3	0.30408	3810	0.15	0.25	0.43	0.17	0.37449
	Old			3810	0.03	0.06	0.22	0.69	
English Language	New	0	0.00084	118,085	0.12	0.20	0.36	0.32	0.03060
	Old			118,087	0.12	0.20	0.36	0.32	
	New	0.25	0.25103	87,085	0.16	0.27	0.48	0.09	0.23591
	Old			87,085	0.04	0.04	0.49	0.43	
	New	0.50	0.49514	54,085	0.26	0.44	0.29	0.01	0.44059
	Old			54,085	0.01	0.01	0.29	0.69	
	New	0.75	0.75046	16,335	0.85	0.09	0.03	0.03	0.72279
	Old			16,335	0.03	0.03	0.09	0.85	

Note. Obs ES$_1$ based on Total N. Obs ES$_2$ based on sample of 1500.

Table 3.11 Equating Conditions

Comparison of SG and CINEG Design Equating Results

Exams		Equating Methods	Criterion
Chemistry French Physics B English	X	Direct Equipercentile IRT True Score (SG) IRT Observed Score (SG) Chained Equipercentile Frequency Estimation IRT True Score (CINEG) IRT Observed Score (CINEG)	Consistency of equating results

Total Number of Equating Relationships: 28 (4 Exams X 7 Equating Methods)

Population Invariance

Exams		Equating Methods		Subgroups	Criterion
Chemistry French Physics B	X	Direct Equipercentile IRT True Score (SG) IRT Observed Score (SG)	X	PED L1 & L2 PED L3 PED L4 Total Group	Total group equating relationship for a given equating method
English			X	PED L1 PED L2 PED L3 PED L4 Total Group	

Total Number of Equating Relationships: 51 (3 Exams X 3 Equating Methods X 4 Subgroups +
1 Exam X 3 Equating Methods X 5 Subgroups)

Comparison of Equating Results with Group Differences (M_0)

Exams		Equating Methods		ES	Criterion
Chemistry French Physics B	X	Chained Equipercentile Frequency Estimation IRT True Score (CINEG) IRT Observed Score (CINEG)	X	0 0.1 0.2 0.3	Equating relationship for a given equating method where ES = 0
English			X	0 0.25 0.5 0.75	

Total Number of Equating Relationships: 64 (4 Exams X 4 Equating Methods X 4 ESs)

Comparison of Equating Results for Matched Samples (M_1–M_3)

Exams		Equating Methods		ES		Matching Method	Criterion
Chemistry French Physics B	X	Chained Equipercentile Frequency Estimation IRT True Score (CINEG) IRT Observed Score (CINEG)	X	0.1 0.2 0.3	X	M_1 M_2 M_3	Equating relationship for a given equating method where ES = 0
English			X	0.25 0.5 0.75			

Total Number of Equating Relationships: 144 (4 Exams X 4 Equating Methods X 3 ESs X 3 Matching Methods)

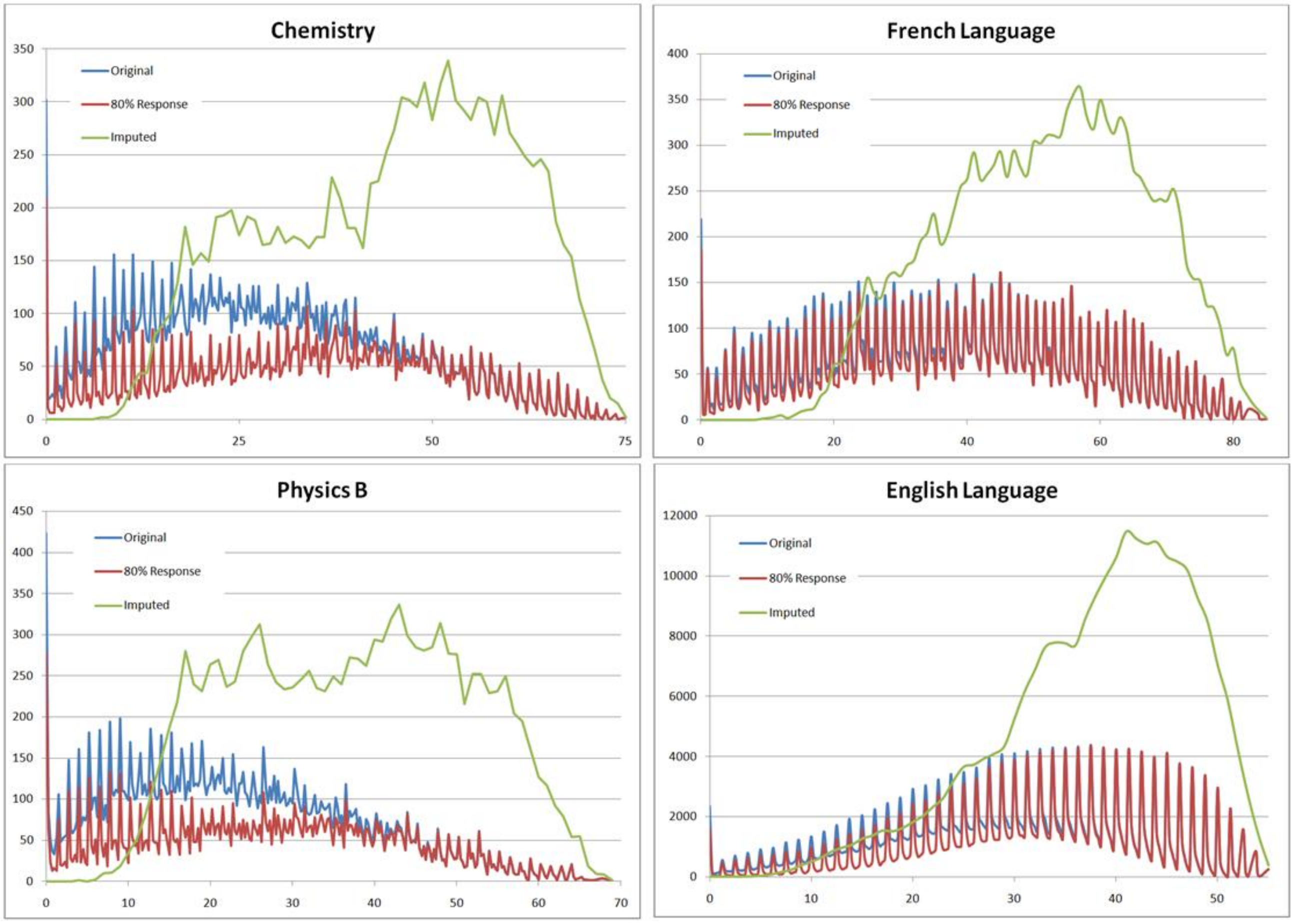


Figure 3.1 Unweighted MC frequency distributions.

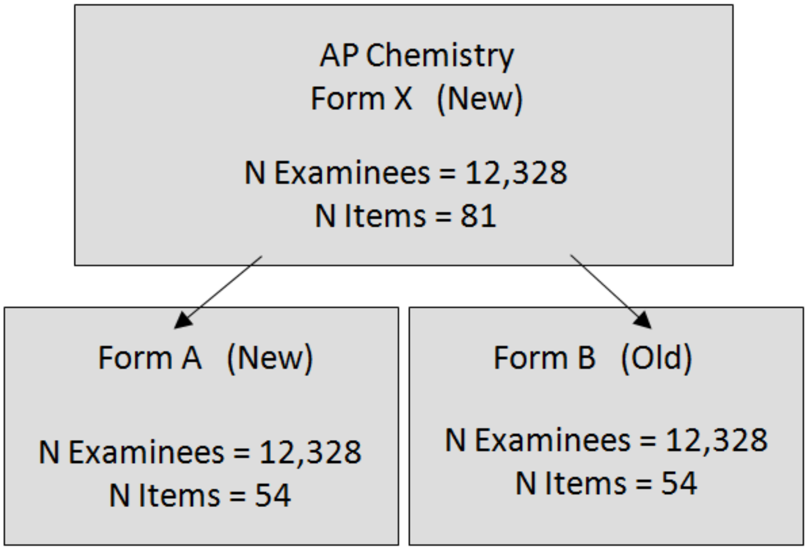


Figure 3.2 Creating Forms A and B from Form X of AP Chemistry.

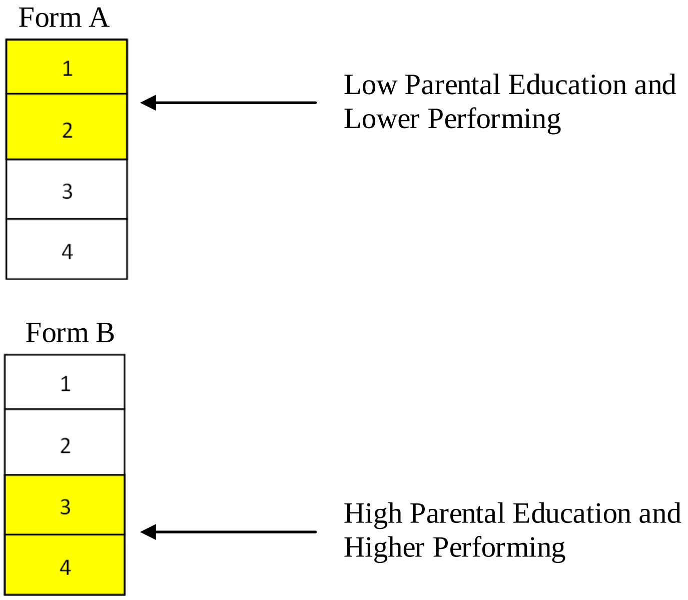

Form A
1
2
3
4
Low Parental Education and Lower Performing
Form B
1
2
3
4
High Parental Education and Higher Performing

Figure 3.3 Creating pseudo groups with varying performance differences.

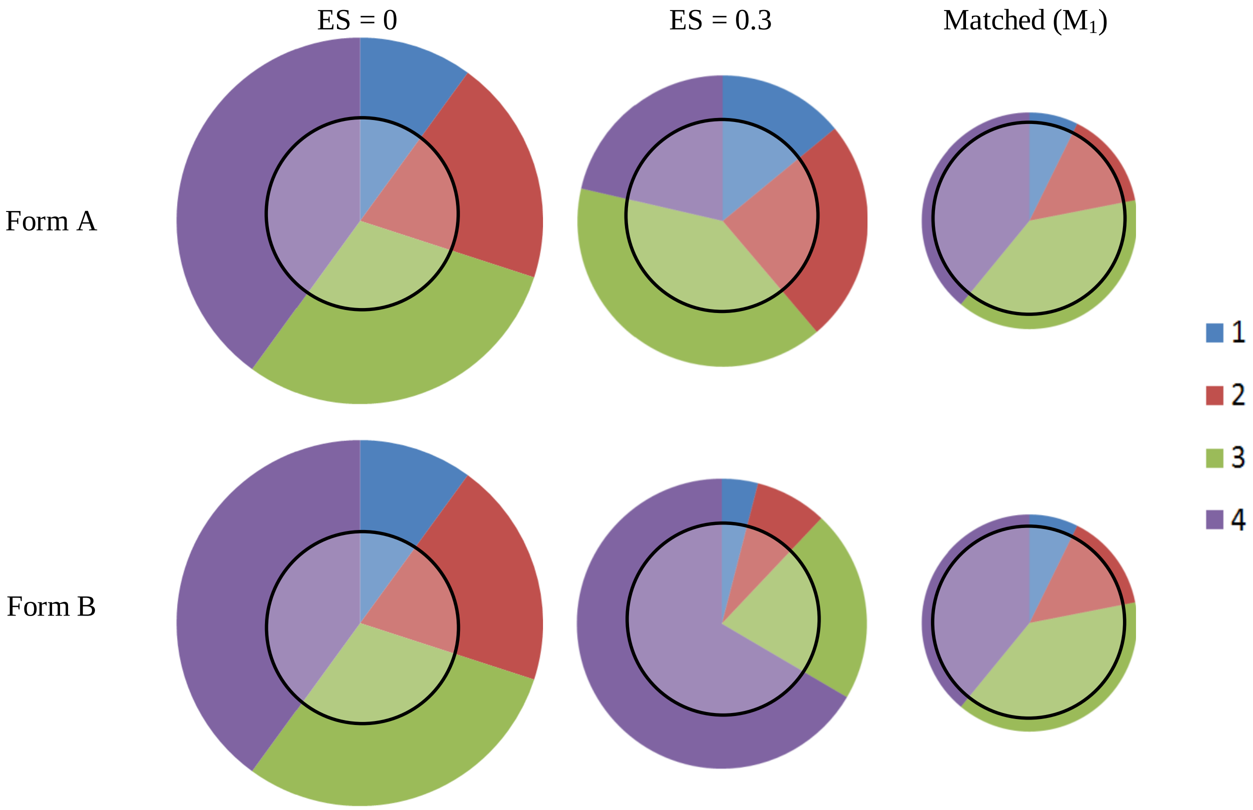


Figure 3.4 AP Chemistry Form A and B Samples.

CHAPTER 4

RESULTS

This chapter contains seven sections. In the first section, the results of SG and CINEG equating methods are compared for equivalent old and new form groups, as a check on the reasonableness of the output from *Equating Recipes*. In the second section, the results of a set of population invariance studies are described as they pertain to Research Question 1. Research Questions 2 and 3 are addressed in the third and fourth sections. A fifth section provides a description of the results for the matching methods considered in this dissertation. Finally, the sixth and seventh sections provide a description of the results for Research Questions 4 and 5.

Comparison of Results using SG and CINEG Methods

SG equating results using the equipercentile, IRT true score, and IRT observed score equating methods were compared to CINEG equating results using the frequency estimation, chained equipercentile, IRT true score, and IRT observed score equating methods. For these comparisons, a stratified random sample of 1500 examinees was selected from each of the four AP exams. Form A and Form B composite scores and common item scores were calculated for all of the examinees. For the SG design, equating was conducted using only the composite scores. For the CINEG design, the common item scores were also used for equating even though the “old” and “new” form groups were the same examinees, and the common item scores were therefore identical for Forms A and B. The use of CINEG equating methods with equivalent groups is not necessary, but was used to ensure that the CINEG equating methods provide identical results to the SG equating methods when the old and new form groups are the same. This comparison was conducted solely for the purpose of checking the reasonableness of the output provided by *Equating Recipes*.

As discussed later, the CINEG frequency estimation and chained equipercentile equatings were smoothed using cubic spline postsmoothing using bootstrap SEs and an *S*-value of 0.1. Although an *S*-value of 0.1 was not necessarily optimal for the SG equatings, smoothed results based on *S*=0.1 appeared to be reasonable, and provided consistency with the *S*-value chosen for the CINEG methods. Figure 4.1 provides an example of the smoothed and unsmoothed equivalents, along with plus and minus one SE for the unsmoothed equivalents for Chemistry for both frequency estimation (CINEG) and equipercentile (SG) equating methods. Note that the equivalents shown in Figure 4.1 are the old form (B) equivalents minus the new form (A) score. Equivalents are provided only for the first through the 99th percentiles because beyond this region, there is very little data with which to estimate the equating relationships.

The smoothed equivalents met the criteria described in Chapter 3: they were smoother than the original unsmoothed equivalents but did not exceed the SE bands within the region of the score range in which examinees scored. The SEs for the SG equatings were much smaller than the SEs for the CINEG equatings. The SEs differed because of the way the bootstrapping procedure was carried out with the CINEG equating methods. With equivalent groups, the SG SEs are the appropriate SEs. However, because all subsequent analyses using nonequivalent groups involve the CINEG bootstrapping method, the larger SEs are provided in Figure 4.1 for comparison. Note that the unsmoothed equivalents in the top and bottom plots of Figure 4.1 are exactly the same. When groups are equivalent, the unsmoothed frequency estimation and unsmoothed direct equipercentile equating results are identical.

When the old and new form groups are the same, as in the SG design, the common item ES is exactly zero. Therefore, the CINEG equating methods adjust only for form differences, not group differences, just like SG equating methods. It was expected that equating relationships for the CINEG and SG equating methods would provide identical results in this situation. The results for the CINEG equating methods are compared to the counterpart SG equating results in Figures 4.2-4.5. For example, IRT true score equating results for the CINEG design are compared to IRT true score equating results for the SG design, and frequency estimation (CINEG) results are compared to equipercentile (SG) equating results.

IRT equating results for the SG and CINEG designs are provided for each exam in the top two plots of Figures 4.2-4.5. For Chemistry and Physics B, the IRT equating results were very similar across the score scale (see Figures 4.2 and 4.5). For English and French Language (Figures 4.3 and 4.4 respectively), the IRT SG and CINEG results were different at the low end of the score range. The equating relationships were not identical because item parameters for Forms A and B were estimated simultaneously for the SG design but separately for the CINEG design. When the simultaneously estimated item parameters were used with the CINEG IRT equating methods, the results were identical to those obtained with the SG IRT equating methods.

The SG and CINEG equivalents for the unsmoothed traditional equating methods were identical, as expected, for all exams (see the middle plots in Figures 4.2-4.5). The SG and CINEG smoothed equivalents for the traditional methods were very similar, especially for Chemistry and Physics B (see the bottom plots in Figures 4.2-4.5). The smoothed results for the two designs differed only because cubic spline postsmoothing involves SEs calculated using a different bootstrap procedure for the SG and CINEG equating methods.

Forms A and B were constructed to be as similar as possible in terms of content and difficulty (see Chapter 3 for a description of how the forms were created). Comparing the closeness of the equating relationship to the vertical axis value of zero indicates that Chemistry had the least similar old and new forms and Physics B had the most similar forms.

The first four moments of the old form equivalent of new form score distributions are provided in Table 4.1 for the postsmoothed traditional equating methods and Table

4.2 for the IRT equating methods. In general, the moments were very similar across all seven equating methods. The postsmoothed traditional equating moments (equipercentile, frequency estimation, and chained equipercentile) were somewhat more similar to each other than to the IRT methods. Likewise, the SG and CINEG IRT equating results tended to be more similar to each other, than to the traditional equating moments. The CINEG IRT means tended to be higher than the SG IRT means, especially for English and French Language. The moments for English and French Language were more similar for the postsmoothed traditional methods than for the SG and CINEG IRT methods. However, all differences between moments and equating relationships were small, as expected. The results obtained with *Equating Recipes* for the CINEG and SG designs were consistent, as expected. The comparison of SG and CINEG results provides a baseline for all subsequent analyses.

Research Question 1

To what extent do equating results appear invariant for populations of examinees with different levels of parental education?

Parental education was categorized into four levels with Level 1 being the lowest level of parental education and Level 4 being the highest level of parental education. To assess the population invariance of equating results based on subgroups of examinees with different levels of parental education, equatings were conducted for samples of 1500 examinees for each of the four categories. For AP Chemistry, French Language, and Physics B, a stratified sample of 1500 examinees was drawn from parental education Levels 1 and 2 combined because neither category contained 1500 examinees alone. For English Language, 1500 examinees were drawn from each level. As a criterion equating relationship, a stratified random sample of 1500 was taken across all four levels of parental education. The stratified sampling was based on the actual proportions of examinees in each parental education level for each exam. Table 3.7 provides the percentage of examinees in each parental education category for the four exams. For example, of the 1500 examinees sampled for the Chemistry criterion equating, approximately 9% were from parental education Level 1, approximately 15% were from parental education Level 2, approximately 33% were from parental education Level 3, and approximately 40% were from parental education Level 4. Note that the actual percentages were slightly higher because examinees with no parental education response were not included in the sampling process. The criterion equating sample is referred to as the total group because it contained examinees from all four levels of parental education.

Population invariance analyses were carried out using the SG design. All examinees had a Form A and a Form B composite score which was used to calculate the equating relationship between the two forms. Three SG equating methods were compared: postsmoothed equipercentile, IRT true score, and IRT observed score. Results were compared in terms of resulting equating relationships, equated score moments, REMSD, and classification consistency.

Comparison of Equating Relationships

Equating results for the four AP Exams, three equating methods, and three or four subgroups are displayed in Figures 4.6-4.9. In each figure, the colored lines in the top plots provide a comparison of the equating results for each subgroup in terms of the difference between the old form equivalents for the subgroup and the old form equivalents for the total group for the IRT true score equating method (left) and the IRT observed score equating method (right). A similar plot is provided in the bottom left for the smoothed equipercentile equating method. The vertical axis value of zero in these three plots is the criterion; that is, 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.

Black lines indicate plus and minus two SEs calculated using 1000 bootstraps for the total group unsmoothed equipercentile equating. The SEs do not apply directly to the IRT equating methods but are still plotted to provide some indication of how similar the subgroup equating results are to the total group equating results. Subgroup equating relationships within the SE bands are considered within sampling error of the total group equating relationship for at least the equipercentile method.

The fourth plot in the bottom right corner of Figures 4.6-4.9 provides a comparison of the total group equating results for the three SG equating methods. Again, the colored lines represent the equating relationships and the black lines represent plus and minus two SEs based on the total group equipercentile equating. In the bottom right plot, the vertical axis value of zero is not the criterion; it is the difference between the old form equivalents and the new form scores. The distance the equating relationships are from the vertical axis value of zero provides an indication of how different the new and old forms are in terms of difficulty. The criterion in the fourth plot is the consistency of equating results across methods.

Note that the bottom right plots in Figures 4.6-4.9 have a different vertical axis scale than the other three plots because the values being compared are different. In the bottom right plot the equivalents are graphed in terms of the old form equivalent minus the new form score. In the top plots and bottom left plot in Figures 4.6-4.9, the vertical scale represents the difference between the old form equivalents for the subgroup and total group equating. In all four plots, equating relationships are only graphed for the first through 99th percentiles because beyond these ranges there is very little data with which to estimate the equating relationships.

A comparison of Figures 4.6-4.9 indicates that the results of the population invariance analyses were quite consistent across exams. For the IRT methods, the subgroup results were within two SEs of the criterion for the majority of the score scale. The subgroup that tended to have the largest deviation from the total group was the subgroup with the lowest level of parental education. The postsmoothed equipercentile results were much less smooth than the IRT results. Increasing the smoothing parameter would have made the postsmoothed results less bumpy but may have introduced more bias. Although postsmoothed equipercentile results were plotted in Figures 4.6-4.9, the equipercentile SEs plotted were based on the unsmoothed equipercentile equating because *Equating Recipes* was set up to handle the calculation of bootstrap SEs of unsmoothed equipercentile equating, but not to handle the calculation of bootstrap SEs of smoothed equipercentile equating. The SEs for postsmoothed equipercentile equating would have been slightly smaller and smoother. In general, the equipercentile subgroup results tended to exceed two SEs more noticeably than the IRT methods. Equipercentile equating relationships exceeded two SEs at some score points for all subgroups. Although increasing the smoothing parameter may have kept the equipercentile equating results within plus or minus two SEs for more score points, the equipercentile equating results for all subgroups were within two SEs of the criterion for the majority of the score scale. A comparison of the equating methods for the total group indicated that all three SG equating relationships were very similar, and for the most part, within two SEs of the total group equipercentile method.

Although Forms A and B were constructed to be as similar as possible in terms of content and difficulty, a comparison of the closeness of the equating relationships to the vertical axis value of zero in the bottom right plots across exams indicated that Form A and B were very similar for Physics B, fairly similar for English and French Language, and less similar for Chemistry. The relative smallness of equipercentile SEs for Physics B may have been caused by the relatively small score scale (0-91) compared to the score scales for the other exams. Chemistry, with the next smallest score scale range, had a maximum of 190 points.

Moments

The first four moments are provided in Tables 4.3-4.6 for old form equivalents based on unsmoothed and postsmoothed equipercentile, IRT true score, and IRT observed score equatings, and for both the total group and all subgroups. Similar patterns were found across exams. First, the means increased as parental education level increased. The total group mean was typically slightly smaller than the mean for the Level 3 subgroup. The standard deviation was often largest for the combined Levels 1 and 2 and smallest for Level 4, although this was not always the case. As parental education level increased, the skewness often shifted from slightly negative or positive, to more negatively skewed. Kurtosis (centered at 3) ranged from slightly platykurtic for Physics B and Chemistry, to slightly leptokurtic for Levels 3 and 4 of English Language.

Across exams, the unsmoothed and postsmoothed equipercentile moments were very similar within each level of parental education. The IRT methods tended to have more similar moments to one another, but the IRT moments were also very similar to the traditional moments.

REMSD

REMSD values, quantifying the difference between each criterion and comparison equating relationship, are provided for all four exams and all three SG equating methods in Table 4.7. To facilitate interpretation of the REMSD values, often the standardized difference that matters (SDTM) is used. SDTM is defined as a half of a score point divided by the old form standard deviation for the total group. The total group old form standard deviations for each exam were provided in Table 3.4 (see COMP SD for Form B). The SDTM for each test was provided in the last column of Table 4.7. None of the equipercentile REMSD values was less than the SDTM, even when smoothing was used. However, many of the IRT true and observed score REMSD values were less than the SDTM. Apart from Physics B, the largest IRT REMSD values (and those that exceeded the STDM) were for the lowest categories of parental education. These findings suggest that 1) the equating relationships based on subgroups that make up the largest proportion of the total group (Levels 3 and 4) are more likely to be closer to the total group equating relationship, and 2) IRT equating results appear more population invariant than the traditional equipercentile equating results.

Classification Consistency

Four cut scores were found for each classification consistency calculation as follows:

1. Operational AP cut scores used with the full-length formula scored noninteger weighted AP Exams.
2. Form B cut scores 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.
3. Form A cut scores for the total group based on the total group equating relationship between Form A and Form B. Cut scores were found for each equating method.
4. Form A cut scores based on each subgroup equating relationship between Form A and Form B. Cut scores were found for each combination of equating method and subgroup.

The operational AP cut scores were provided in documentation from the College Board; the other three cut scores were calculated in this study. Table 4.8 provides AP grades, operational cut scores and cumulative percentages, and Form B cut scores and cumulative percentages for each exam. The Form B cut scores were higher for Chemistry, English, and French Exams because of the higher number correct scores compared to formula scores and higher integer section weights compared to the operational noninteger section weights. For Physics B, the cut scores for Form B were smaller than the operational cut scores because the integer section weights were smaller than the noninteger operational weights. The cumulative percentages (% Below) for the operational and Form B cuts were very similar, as expected. Because the four exams have different score ranges, the cut scores are not directly comparable. However, the cumulative percentages indicate that operationally English Language had the lowest percentage of examinees receiving an AP score of 1 (11.28%), and the second lowest percentage of examinees receiving an AP score of 5 (100-91.31=8.69%). Operationally French Language had the lowest proportion of examinees receiving an AP score of 5, and the second highest percentage of examinees receiving an AP score of 1 (23.51%).

Tables 4.9-4.12 provide the Form A cut scores for the total group and all subgroups. Although the cut scores were fairly consistent across subgroup and method, small differences were found for all exams. Table 4.13 provides the classification consistency values for each exam, equating method, and subgroup. Classification consistency is 100% when the cut scores for the total group and the cut scores for the subgroup are identical. All methods and subgroups had at least 95% classification consistency with the total group classification. The IRT equating methods tended to have higher classification consistency than the equipercentile equating methods. The criterion classification was based on the new form cut scores found for the total group for a given equating method. Therefore IRT true score classification consistency was based only on the total and subgroup cut scores found using the IRT true score equating method. There did not appear to be any relationship between the classification consistency value and the subgroup. Classification consistency values tended to be higher for Chemistry and English Language across all subgroups and equating methods.

Research Question 2

What is the impact of group differences on equating results?

Research has shown that commonly used equating methods like frequency estimation, chained equipercentile, IRT true score, and IRT observed score, provide very similar results when old and new form groups are fairly similar (Kolen, 1990). However, there is some indication that large group differences may cause equating methods to provide divergent and inaccurate results (e.g., Eignor, Stocking & Cook, 1990; Walker, Allspach, & Liu, 2004). In order to investigate the impact of group differences on equating results, the ES difference in common item performance between the old and new form groups was manipulated to systematically differ. Examinees from higher levels of parental education were sampled in greater proportions for the old form group, and examinees from the lower levels of parental education were sampled in greater proportions for the new form group. The sampling from parental education levels resulted in a common item standardized mean difference (or ES) between the old and new form groups of approximately 0, 0.1, 0.2, and 0.3 for AP Chemistry, French Language, and Physics B, with the old form group having a higher average performance. For AP English Language, the sampling of examinees from the four levels of parental education was modified so that the ES between old and new form groups was approximately 0, 0.25, 0.50, and 0.75. The larger sample size for English Language allowed for the sampling (and later, matching), of old and new form groups with more extreme differences than was possible with the other three exams. Note that the sampling procedure resulted in observed ES values that were not exactly equal to the target ES values (see Table 3.10 Obs ES_2). However, in all further discussion, the ESs are referred to by the target values (i.e., 0, 0.1, 0.2, 0.3, 0.25, 0.50, and 0.75). Results for the four ES levels were compared in terms of the equating relationships, equated score moments, REMSD, and classification consistency.

Comparison of Equating Relationships

For each exam, the four CINEG equating methods, postsmoothed frequency estimation, postsmoothed chained equipercentile, IRT true score, and IRT observed score, were compared for each of four ESs. The hypothesis was that as ES increased, the equating results would become less similar across methods.

Figure 4.10 provides a comparison of the equating relationships for different ESs for each of the four equating methods for AP Chemistry. The criterion equating relationship (ES=0) is represented by the vertical axis value of zero. The colored lines represent the comparison equating relationships. The black lines represent plus and minus two bootstrap SEs from the zero value on the vertical axis based on the chained equipercentile method for the criterion equating (ES=0). Although frequency estimation bootstrap SEs were also calculated, they were similar (albeit slightly smaller) than the chained SEs. To simplify comparisons, only chained SEs were plotted.

The top two plots in Figure 4.10 provide the Chemistry equating results for the two IRT methods. The equating relationship for each comparison ES (0.1, 0.2, and 0.3), was compared to the equating relationship for the criterion (ES= 0) by taking the difference between the equivalents. The closer the comparison equating relationship is to the vertical axis value of 0, the closer it is to the criterion equating relationship. Surprisingly, for Chemistry the equating relationship for the smallest ES (0.1) was the farthest from the criterion equating relationship for the IRT methods. The bottom two plots provide the equating relationships for the two traditional methods. The equating relationships for all three comparison ESs exceeded plus or minus two SEs at some score points. Equating results for ES=0.1 were the most divergent from the criterion which was unexpected but most likely due to sampling error. In general, the equating relationships for the postsmoothed traditional methods were less smooth than the equating relationships for the IRT methods. Increasing the smoothing parameter for the traditional methods would have resulted in smoother equating relationships but may have introduced additional bias into the equating relationships.

Figure 4.11 provides a comparison of the four equating methods at each of the four ES levels for AP Chemistry. As in Figure 4.10, the colored lines represent the equating relationships and the black lines represent plus and minus two bootstrap SEs based on the chained equipercentile method for the ES shown.

The top left plot provides the equating relationships for the four equating methods when the ES was approximately 0. All methods provided very similar results which were within plus or minus two SEs. The top right plot provides the equating relationships for the four methods when the ES was approximately 0.1. There is a noticeable difference in the equating relationships for the four methods. The IRT equating methods both produced nearly identical results throughout the score range but the IRT and traditional equating results were quite different and the IRT methods exceeded the SE bands at some score points. The results for the two traditional methods were fairly close at ES=0.1. The bottom two plots provide the equating relationships for ESs of 0.2 and 0.3. Although differences between the IRT and traditional equating results did not appear as large as they did for an ES of 0.1, the results for the two traditional methods appeared to become less similar as the ES increased. For Chemistry, the IRT true and observed score equating results were nearly identical even for the largest ES. Despite noticeable differences between the IRT and traditional equating results, for all four ESs, the traditional and IRT equating results were within plus or minus two SEs for the majority of the score range.

Figures 4.12-4.13 provide English Language ES comparisons for each equating method. The group differences for English Language were much more extreme than the differences for the other exams, which was possible because of the much larger initial sample size for the English Language Exam compared to the other exams. A comparison of the four plots in Figure 4.12 indicates that equating relationships for all ES values deviated from the criterion equating relationship by more than two SEs for a large region of the scale. 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 4.13 provides a comparison of the four equating methods at each ES for English Language. 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 sparseness of data at the low end of the score scale contributed to the large SEs and the observed differences in the results for the different equating methods. At higher scores, where more examinees scored, the equating relationships were all very close for ES=0. The IRT observed and true score methods were very similar to one another for all ESs except at extreme scores. A comparison of the equating relationships across the four plots indicates that the IRT equating relationships changed substantially from ES=0 to ES=0.75, as was indicated in Figure 4.12. As ES increased, the results for the traditional methods became less similar to one another and less similar to the IRT equating results.

Comparisons of the equating relationships at each ES are provided in Figure 4.14 for French Language. The IRT results for all ES levels were within plus or minus two SEs of the criterion equating relationship. The frequency estimation equating results for ES=0.2 and 0.3 exceeded two SEs for several score points. The chained equating results for all ES levels were within two SEs for the majority of the score scale.

A comparison of the results for the four equating methods at each ES is provided in Figure 4.15 for French Language. At ES=0 the equating results differed noticeably at the low end of the scale. However, for the range of scores where most examinees scored, the methods all provided very similar results. As with the other two exams, the IRT methods provided nearly identical results at each ES. The traditional methods appeared to become more divergent from each other and from the IRT methods as ES increased. However, even at ES=0.3, all four equating methods were within plus or minus two SEs of one another.

Figures 4.16 and 4.17 provide a comparison of equating methods and ESs for Physics B. Note that the vertical scale was shorter for Physics B compared to the vertical scales used with the other exams. Smaller differences among the equating relationships were due in part to the smaller score range and standard deviation for Physics B. For the IRT methods, the 0.1 ES provided equating results most divergent from the criterion equating results. For the traditional methods, 0.1 and 0.3 ESs resulted in equating relationships that deviated most from the criterion. For the IRT and chained methods, the equating relationships for all ES levels were within plus or minus two SEs for the majority of the score scale. For the frequency estimation method, the equating relationship for ES=0.3 deviated from the criterion by more than two SEs for a large number of score points. A comparison of equating results across methods indicated that the IRT and traditional equating results were very similar, even at the largest ES. The four methods provided results that were within plus or minus two SEs even at ES=0.3 (see Figure 4.17).

In general, differences in the traditional equating results increased as ES increased for all four exams. However, even for very large ESs the equating results for the two IRT methods did not appear to diverge except at the very extremes of the score scale where there was little data. However, increases in ES resulted in increases in the divergence of comparison and criterion equating results only for the very extreme English Language ESs. With ESs ranging only from 0 to 0.3, there was no consistent pattern between the magnitude of the ES and the extent to which a comparison equating relationship deviated from the criterion relationship. In fact, most of the equating relationships for ESs ranging from 0 to 0.3 were within plus or minus two SEs from one another.

Moments

The first four moments are provided for each equating method and ES combination for the four exams in Tables 4.14-4.17. Moments are provided for unsmoothed frequency estimation (UFE), smoothed frequency estimation (SFE), unsmoothed chained equipercentile (UCE), smoothed chained equipercentile (SCE), IRT true score, and IRT observed score equating methods. The same general pattern of results was found for each exam. Because the old form group was sampled to be higher performing than the new form group, the old form equivalent means typically decreased for all equating methods as the ES increased. At an ES of 0, all methods produced similar equated score moments. The smoothed and unsmoothed moments were very similar for the frequency estimation method at each ES and the same was true for smoothed and unsmoothed chained equipercentile moments. Also, the two IRT methods had very similar moments at each ES, even 0.75. However, as the ES increased, the means became less similar for the traditional and IRT methods. For all exams, as ES increased, the frequency estimation method means became increasingly higher than the IRT method means. The means for the chained method tended to fall in between the frequency estimation and IRT means. The difference between the chained and IRT means also tended to increase as ES increased.

REMSD

REMSD and SDTM values are provided for all exams, ESs, and equating methods in Table 4.18. The values for frequency estimation and chained equipercentile are based on the smoothed equivalents. For Chemistry, French, and Physics B, there was no obvious trend in the magnitude of the ES and the magnitude of the REMSD value. This was not surprising, given the graphical comparisons of the equating difference plots described previously. For English Language however, the REMSD values increased as the ES increased, as expected. The relationship between ES and equating accuracy, as measured by the REMSD statistic, may only become apparent with more extreme ESs. The only exam with REMSD values lower than the SDTM was Physics B. For the two traditional methods, the REMSD for an ES of 0.2 was lower than the SDTM; for the two IRT methods, the REMSD for an ES of 0.3 was smaller than the SDTM. English Language had the highest REMSD values, as expected, given the larger ESs.

A comparison of Tables 4.7 and 4.18 indicates that for all exams besides Physics B, 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 based on any of the parental education subgroups. 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

Form A cut scores based on each combination of ES and equating method are provided in Tables 4.19-4.22 for the four exams. Compared to the SG cut scores (see Tables 4.9-4.12), there appears to be considerably more variability in the CINEG cut scores across methods and ESs. English Language appeared to have the most variability which was expected given the more extreme ESs. Physics B appeared to have the least variability but it also had less than half of the number of score points compared to the other exams.

Classification consistency values are provided in Table 4.23, where ES=0 is the criterion equating relationship, and each equating method is compared only to itself. The magnitude of the classification consistency values did not appear to correspond in a predictable way to the magnitude of the ES for Chemistry, French Language, or Physics B. For English Language, the highest classification consistency was found for the lowest ES (0.25) for all four equating methods. English Language classification consistency was particularly low for the frequency estimation method, and the values decreased as the ES increased. The IRT classification consistency values were higher across all three ESs, compared to the traditional method values. Across all exams, classification consistency values ranged from 70.11 to 100%. Physics B had the highest average classification consistency, and English Language had the lowest average.

In Table 4.24, classification consistency values are provided using IRT true score equating results as the criterion. For example, in Table 4.24, the value of 93.67 in the bottom left of the Chemistry section indicates that the frequency estimation method, when used to equate old and new form groups that differed by 0.3SD units, resulted in 93.67% of examinees receiving the same grade that they would have received based on the IRT true score equating method used with the old and new form groups that differed by 0.3SD units. There were three main patterns found across all four AP Exams in Table 4.24: 1) IRT methods provided very similar results in terms of AP grades (i.e., new form cut scores), 2) the traditional equating methods tended to have lower classification consistency than IRT observed score equating with IRT true score equating, and 3) classification consistency was not clearly related to ES except for very large ESs with the frequency estimation method. Note that classification consistency in Table 4.24 does not indicate the accuracy of the equating methods; the values indicate the consistency of examinee classifications for the four equating methods. These findings parallel those found with the difference plot comparisons described above (see Figures 4.10-4.17).

Research Question 3

What is the impact of group differences on the degree to which equating assumptions are met?

Forms A and B were created by sampling items out of one operational form. Therefore, even with the CINEG design, scores could be calculated for both forms for each examinee. Using the scores on both forms, frequency estimation assumptions and chained equipercentile assumptions were assessed directly. IRT assumptions were also assessed using scores on Forms A and B. The equating assumptions were assessed at each ES using samples of 1500 examinees for both the old and new form groups.

Frequency Estimation

The statistical assumptions for the frequency estimation method are 1) that the conditional distribution of Form A scores given common item scores (*V*) is the same in both old and new form groups, and 2) that the conditional distribution of Form B scores given *V* is the same in both old and new form groups. To assess these assumptions, the largest difference in cumulative frequencies for the old and new form groups was calculated for each of *V* conditional distributions for Form A composite scores. The number of old and new form examinees at each *V* was then used to calculate the weighted average of maximum differences across all *V*. The same procedure was used to calculate a weighted average of maximum differences across all *V* for conditional Form B composite scores. Table 4.25 provides the weighted averages for all exams and ESs. The third column provides the weighted maximum differences for Form A conditional distributions, and the last column provides the corresponding values for Form B conditional distributions. Both columns provide very similar numbers, indicating that the degree to which the old and new form groups differed in terms of their conditional distributions on Form A was similar to the degree to which they differed in terms of their conditional distributions on Form B. The pattern across exams was clear: as ES increased, the weighted maximum differences increased for both forms. In other words, as old and new form groups become less similar, the degree to which frequency estimation assumptions hold decreases.

The values in Table 4.25 were compared to the REMSD values for the frequency estimation method in Table 4.18. It was expected that higher weighted maximum differences (assumptions holding less well) in Table 4.25 would lead to higher REMSD values (less accurate equating results) for the frequency estimation method.

However, only English Language showed a direct correspondence between the degree to which the frequency estimation assumptions held, and the accuracy of the equating results.

Chained Equipercentile

The statistical assumptions for the chained equipercentile equating method are 1) that the equating relationship from Form A composite scores to *V* is the same in both old and new form groups, and 2) that the equating relationship from *V* to the composite scores on Form B is the same in both old and new form groups. To assess these assumptions, the *A* to *V* and *V* to *B* SG equipercentile equating relationships were calculated for the new group and the old group. The difference between each equating relationship for the two groups was quantified using the REMSD statistic. Higher REMSD values indicated that the chained assumptions did not hold as well. Table 4.26 provides the REMSD values for the *A* to *V* (column 3) and *V* to *B* (column 4) equating relationships. The numbers in columns 3 and 4 are fairly similar indicating that the impact of different ESs on the similarity of *A* to *V* equating results for old and new form groups is similar to the impact of different ESs on the similarity of *V* to *B* equating results for the two groups. REMSD values increased consistently as ES increased for English Language, but for the other exams the ES and REMSD values did not appear related. A comparison of the REMSD values in Table 4.26 to the chained REMSD values in Table 4.18 indicated that the degree to which the chained assumptions held was directly related to the closeness of the comparison and criterion equating relationships. In other words, it appears that comparison groups (i.e., ES>0) for which the chained assumptions held to a closer degree (i.e., smaller REMSD values in Table 4.26), produced chained equating results that were closer to the criterion equating relationship (ES=0). This finding highlights the fact that equating results are more accurate when equating method assumptions hold. The less well the assumptions hold, the less accurate the equating results are likely to be.

IRT

A major assumption for the two unidimensional IRT equating methods is that the two forms measure the same unidimensional construct. Several methods were used to assess the dimensionality of the four exams, including the MC-FR correlations, principal component analysis, dimensionality software including PolyDIMTEST and PolyDETECT, and finally, a comparison of item parameter estimates.

MC-FR Reliability and Correlations

One possible source of multidimensionality with mixed-format exams is the different item types. MC and FR items are often included in an exam to measure somewhat different aspects of a content domain. The lower the MC-FR correlation, the more likely that each item type measures a different construct which would indicate that the assumption of unidimensionally does not hold for the exam. Coefficient alpha was calculated for both the MC and FR sections to provide a reliability estimate used to disattenuate the observed correlations. The reliability estimates, observed correlations, and disattenuated correlations are provided in Table 4.27 for each exam, form, and ES. The correlations and reliability estimates were fairly similar across ES samples and forms. The MC reliability was fairly high for all four exams (the lowest being English Language). The reliability of the FR section was highest for French and Physics B. Observed correlations were above 0.75 for both Chemistry and Physics B. The disattenuated correlations were also highest for Chemistry and Physics B. The Chemistry and Physics B Exams were considered most likely to be unidimensional so the high MC-FR correlation was not surprising. The language exams had much lower disattenuated correlations, especially English Language (average=0.68). These results provided some evidence that the language exams may not meet the unidimensionality assumption.

However, there was no indication that the magnitude of group differences had any impact on the degree to which the unidimensionality assumption held.

Principal Component Analysis

Principal component analysis was conducted using polychoric correlations calculated for all MC and FR items in the four AP Exams. The process was replicated for each ES sample for both Form A and B. Tables 4.28-4.31 provide the eigenvalues and the percentage of variance explained by each of the first five principal components. There were slight differences in the magnitude of the eigenvalues and percentage of variance explained between the forms, the most extreme difference being for Chemistry (see Table 4.28), where the first principal component accounted for approximately 32% of the variance in Form A, but almost 40% of the variance in Form B. Across all exams, forms, and ESs, the first principal component accounted for less than 40% of the variance. Although a substantial proportion of the variance remained, there was a sharp decrease in the amount of variance explained by the second principal component. The second principal component accounted for approximately 3-4% of the remaining variance, but the ratio of the second principal component to the third was relatively small for all four exams. 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 ES=0 condition. The scree plots for other ESs were nearly identical. Figure 4.18 provides the scree plots for Form A and Form B of Chemistry (ES=0). The larger plots provide the eigenvalues for all principal components. The smaller plots, in the top right corners of large plots, provide 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 both forms of Chemistry. However, there also appeared to be a break between the second and third principal components in the smaller plots. These results indicate that Chemistry may have one dominant dimension and one smaller secondary dimension.

Figure 4.19 provides similar results for English Language. Again, the magnitude of the first eigenvalue is very large compared to the rest. In the large plot there is a clear break between the first and second principal components, but the smaller plot indicates the possible presence of another smaller dimension. In Figure 4.20, the larger plots indicate one dominant dimension for French Language. However, a break between the third and fourth principal components in the smaller plots may indicate that French Language has a three dimension solution. Finally, Figure 4.21 provides scree plots for Forms A and B of Physics B. The large plots again show one dominant dimension, but the smaller plots indicate the presence of at least one smaller dimension. These results suggest that all four AP Exams may violate the unidimensionality assumption to some extent.

PolyDIMTEST and PolyDETECT

One program that was considered for use in assessing the dimensionality of the AP Exams was PolyDIMTEST (Li & Stout, 1995). Data from the full-length operational exam forms were used with PolyDIMTEST. The FR items were used as the assessment subtest (AT1). To compare the results to a baseline condition, another replication was run using ten randomly selected items as AT1. It was expected that the p-values for the DIMTEST statistic would be relatively high for the random set of AT1 items because random items are more likely to be similar in dimensionality to the whole exam form. The p-values for an AT1 comprised of only FR items were expected to be lower because of the homogeneity of the items in comparison to the whole exam. Results were compared for samples of 5,000 and 1,000, and with minimum cell sizes ranging from 2 to 100. These conditions were selected to determine how sensitive PolyDIMTEST results are to program specifications. Based on the results obtained for the four AP Exams included in this dissertation, it appears that PolyDIMTEST is very sensitive to the cell size and sample size specifications. For example, when 1000 Chemistry examinees were included in the dimensionality analysis using the AT1 comprised of FR items, a cell size of 20 resulted in a p-value of 0. However, when the cell size was changed to 2, the p-value was 0.8732. In addition, PolyDIMTEST can only handle up to 12,000 examinees. Three randomly equivalent English Language data sets were sampled from approximately 250,000 examinees. For randomly equivalent sample sizes of 5000, and minimum cell sizes of 50, the p-value based on the AT1 comprised of FR times ranged from 0 to 0.9997. The p-values based on the random item AT1 were also variable and sometimes unexpectedly less than 0.05.

The lack of consistent PolyDIMTEST results across randomly equivalent data sets, and for different program specifications, made the usefulness of this program questionable as a tool to assess dimensionality.

Another program used to assess dimensionality was PolyDETECT (Zhang, 2007). Although PolyDETECT can handle mixed-format data, the program is limited in its usefulness by its requirement that FR items not exceed the range of zero to nine score points. English Language and French Language were the only exams in this study that met this requirement. For English Language, a variety of program specifications was tried. Minimum cell sizes of 5 and 50, dimensions of 2, 3, 4, 5, 6, and 10, and permutations of 2 to 40 were considered. In addition, the dimensionality was assessed including all MC and FR items, and including MC items only. Unlike the results found with PolyDIMTEST, changes to the PolyDETECT program parameters—including the number of dimensions, the number of permutations, and the minimum cell sizes—did not appear to greatly impact the results.

For English Language, when including both MC and FR items, the essential number of dimensions suggested by PolyDETECT was two. The MC items clustered together and the FR items clustered together. Extra MC clusters did not replicate across randomly equivalent English Language data sets. However, when including only the MC items, PolyDETECT provided four or five item clusters. In general, the items that clustered together across the two replications were items within the same MC testlet.

For French Language, when FR items were included, PolyDETECT indicated that there were three essential dimensions. Generally, the items that clustered together were the MC listening items, the MC reading items, and the FR items. However, when FR items were not included, PolyDETECT still provided a three or four cluster solution. Again, listening items tended to cluster together. However, three reading clusters were also provided, corresponding roughly to MC reading testlets.

Because assessing the dimensionality of the FR section would have required collapsing FR score categories as high as 17 into 10 or fewer, PolyDETECT was not used with the Chemistry and Physics B FR sections. For both of these exams, the program indicated that the MC sections were essentially unidimensional.

These results indicate that the English Language and French Language Exams may not be unidimensional. Physics B and Chemistry MC sections may be unidimensional, but the dimensionality of the full forms was not determined with PolyDETECT. Moreover, PolyDETECT results were not consistent when using MC items only or when using both MC and FR items. Because of the limitations of both PolyDIMTEST and PolyDETECT, the programs were not used with Forms A and B.

Item Parameter Estimation

MULTILOG software was used to estimate item parameters for the MC items using the 3PL model and to estimate item parameters for FR items using the GRM model. 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. For the FR items, English Language 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. It was hypothesized that if the two sections measured different constructs, item discrimination values would decrease when calculated for the two sections simultaneously. In other words, with a more complex exam, each item would be less representative of the total exam and would be less discriminating than items in a more homogeneous exam.

Scatter plots of item parameters estimated with and without the other section's items are provided in Figures 4.22-4.25 for AP Chemistry. Scatter plots were only included for ES=0 because the results were similar across ESs for Chemistry. For the MC item parameter estimates, inclusion of FR items tended to increase not only the discrimination parameter estimates, but also the *b*- and *c*-parameter estimates when compared to the MC-only item parameter estimates for Form A (see the left plot in Figures 4.22-4.24). For Form B, the item parameter estimates were fairly consistent regardless of whether or not the FR items were included in the estimation process. The *b*-parameter estimates appeared the most stable; the *c*-parameters appeared the least stable.

Figure 4.25 provides scatter plots of the FR item parameters estimated with and without MC items. Again, the left plot corresponds to Form A values, the right to Form B values. Chemistry Forms A and B had four FR items with 10, 16, 11, and 10 categories. The four *a*-parameters were plotted with red circles. For each FR item, there were the number of categories minus one *b*-parameter estimates. These values were plotted with black circles. For Chemistry, the FR *b*-parameters were very close to the diagonal line, indicating that the inclusion of MC items did not greatly impact the FR *b*-parameter estimates. However, the *a*-parameter estimates tended to lie above the diagonal line indicating that the FR discrimination estimates increased when the MC items were not included.

Table 4.32 provides the average item parameters for each form and ES both with and without the simultaneous estimation of the other section's item parameters. Negative values in column five (Ave Diff) indicate that the average values for the parameter estimates were smaller when the FR and MC sections were estimated simultaneously. Positive values indicate that the parameter estimates were larger when the two sections were estimated together. The standard deviations are also provided under both conditions, as well as the difference between the standard deviations (SD Diff). Positive SD Diff values indicate that the item parameters had greater variability when estimated with both sections, than with only one section. For the Chemistry MC item parameter estimates, the Ave Diff values were all positive, indicating that the MC item parameter estimates found when estimating MC and FR items together were higher than the MC item parameter estimates found when estimating parameters for only the MC items. However, a comparison of the Ave Diff values for Forms A and B indicates that the difference between the MC item parameter estimates was largest for Form A. The same conclusion was drawn from the scatter plots in Figures 4.22-4.24. A comparison of the SD Diff values indicated that the variability was larger for the MC *a*-parameter estimates when the two sections were estimated simultaneously. The MC *a*-parameter variability increased only slightly for Form B but more noticeably for Form A. The variability of the MC *b*-parameter estimates was slightly smaller when the MC and FR item parameters were estimated together for Form A. For Form B, the SD Diff values were very small but consistently positive, indicating an increase in *b*-parameter variability when the item parameters were estimated together.

For the FR items, all Ave Diff values were negative for the *a*-parameter estimates, indicating that the discrimination estimates decreased for FR items when estimated with MC items. This pattern was consistent for both forms and was also noted in Figure 4.25. A comparison of the SD Diff values for the *a*-parameter estimates indicates that the FR *a*-parameters were more variable when using simultaneous estimation than FR-only estimation for Form A. For Form B however, the FR-only *a*-parameter estimates were more variable. The *b*-parameter estimates were very similar for FR-only and simultaneous estimations. For both forms, the *b*-parameter estimates were more variable when FR and MC items parameters were estimated together.

Scatter plots are provided for the MC and FR item parameter estimates for English Language in Figures 4.26-4.29. For brevity, the figures only provide results based on an ES of 0. However, the scatter plots were not nearly as consistent across ESs as they were for other exams. In general, the MC item parameters were fairly similar whether they were estimated alone or simultaneously with the FR items. For FR items, shown in Figure 4.29, the *b*-parameter estimates were more extreme when estimated with the MC items. The *b*-parameter estimates ranged from approximately -7 to 7 for simultaneous estimation but only -5 to 5 for FR-only estimation. The more extreme values for the simultaneous condition caused negative *b*-values to be above the diagonal line, and positive *b*-values to be below the diagonal line. The FR *a*-parameters were consistently higher when estimated without MC items.

For English Language, Table 4.33 provides the average and SD of item parameter estimates for the simultaneous and the separate estimation conditions, as well as the difference in the average and SD for the two conditions. As with Chemistry, the Ave Diff column contained all positive values for the MC item parameters, indicating that all MC item parameter estimates increased when estimated simultaneously with the FR items. Also, the Ave Diff values for the FR *a*-parameters were all negative, indicating that the FR item discrimination values decreased when estimated with the MC items. The SD Diff values were mostly negative for the MC and FR *a*-parameters, indicating that simultaneous estimation resulted in less *a*-parameter variability for both MC and FR items. The SD Diff values for the FR *b*-parameters were large and positive, indicating that simultaneous estimation resulted in much more variable *b*-parameter estimates than the *b*-parameter estimates using FR items only. This pattern was also noted in Figure 4.29.

Scatter plots of French Language item parameter estimates are provided in Figures 4.30-4.33. Form A MC *a*-parameters appeared larger when the FR items were not included in the estimation—a pattern only found for Form A of French Language. Form B showed the usual pattern of larger MC *a*-parameter estimates when the FR items were included in the estimation. The Form A MC *b*- and *c*-parameter estimates appeared more similar for the two estimation conditions compared to the Form B MC *b*- and *c*-parameter estimates. Four sets of FR item parameters are plotted in Figure 4.33: *a*- and *b*-parameters for long (L) FR items and *a*- and *b*-parameters for fill-in type short (S) FR items. Forms A and B of French Language each had 15 S FR items and 4 L FR items. The range of the L *b*-parameters was greater than the range of the S *b*-parameters because there was only one *b*-parameter for each S FR item. The range of the S FR *a*-parameters was larger than the range of the L FR *a*-parameters, probably because there were nearly four times as many S FR items. In general, the S FR *b*-parameters appeared least affected by the estimation method (simultaneous or FR-only). The other parameters all appeared larger when only the FR items were estimated.

Table 4.34 provides the average and SD of item parameter estimates for the simultaneous and the separate estimation conditions, as well as the difference in the average and SD for the two conditions, for French Language. Except for the MC *a*-parameters for Form A, all of the Ave Diff values for the MC item parameters were positive, indicating that the inclusion of FR items in the estimation process resulted in higher MC item parameter estimates. Conversely, for FR items, the Ave Diff values were all negative, indicating that the item parameter values decreased when MC items parameters were estimated simultaneously. In addition, negative values for SD Diff for MC *b*-parameters and FR *a*-parameters indicate that there was less variability in these parameter estimates when the parameters for both item types were estimated simultaneously. Positive values for SD Diff for FR *b*-parameters indicates that the variability of FR *b*-parameter estimates was greater when MC and FR items were estimated together.

Finally, scatter plots for Physics B item parameter estimates are provided in Figures 4.34-4.37. The MC item parameter estimates were very close to the diagonal line, and therefore very similar for both estimation conditions. The same was true for FR *b*-parameter estimates. The FR *a*-parameter estimates were consistently above the diagonal line, indicating that the discrimination values were higher when they were estimated using only FR items.

Table 4.35 provides the average and SD of item parameter estimates for the simultaneous and the separate estimation conditions, as well as the difference in the average and SD for the two conditions, for Physics B. As with the other three exams, the MC item parameter estimates were consistently higher when estimated with the FR items. The FR *a*-parameter estimates were also larger when only the FR parameters were estimated. For the most part, the variability of the FR *a*-parameters decreased when FR and MC parameters were estimated together. However, the variability of FR *b*-parameters increased when FR and MC parameters were estimated together.

The original hypothesis, that the discrimination values would decrease for the MC and FR sections that measured somewhat different constructs when the item parameters were estimated simultaneously, was not substantiated. Instead, MC item parameters increased when the MC and FR item parameters were estimated together, FR discrimination values decreased, and the changes in FR *b*-parameter estimates were not consistent across exams. In addition, the two estimation conditions often resulted in a change in the variability of item parameter estimates, most notably for English Language FR *b*-parameter estimates.

Across all exams and forms, there were two consistent patterns between the magnitude of the ES and the average item parameter estimates (see Tables 4.32-4.35):

1. As ES increased, the Form A *b*-parameters for both MC and FR items increased, regardless of whether the items were estimated simultaneously or separately.
2. As ES increased, the Form B *b*-parameters for both MC and FR items decreased, regardless of whether the items were estimated simultaneously or separately.

As ES increased, the Form A group became lower performing, and the Form B group became higher performing. MULTILOG scales Form A and Form B groups to have the same mean (0) and standard deviation (1) despite the performance differences between the two groups. Therefore items look more difficult (higher *b*-values) when the group is lower performing.

Comparison of Equating Results

Equating methods used with the CINEG design have much more stringent statistical assumptions than equating methods used with the SG design, especially for the traditional equating methods. Differences between CINEG and SG equating results provide indirect evidence that the equating assumptions involved with the CINEG equating methods may not hold. In the first section of this chapter, the CINEG and SG results were compared using SG data. The postsmoothed traditional and IRT results were nearly identical. Therefore, the comparison of equating results for comparison equating (ES>0), to the criterion equating relationship (ES=0), provides an indication of how well the CINEG equating assumptions held. These comparisons were made above, to answer Research Question 2 and are not repeated here. However, there was indirect evidence that the assumptions may have been violated for all of the equating methods because as the ES increased, there was a tendency for the equating results to become further away from the criterion equating results. This was most apparent for English Language which had the most extreme ESs.

Matching

The results presented above indicate that large group differences can cause inaccurate equating results and violations of the statistical assumptions involved with the equating methods. Matched samples equating may improve the degree to which equating assumptions hold and the accuracy of the equating results. Three different matching methods were used with the three comparison ES groups. The first matching method matched old and new form groups in terms of the proportion of examinees in each of the four levels of parental education. Because parental education was the selection variable, the first matching method provided the best case scenario for matched samples equating. However, operationally, the selection variable is usually unknown. Often several background variables are available for examinees and matching is done using a combination of the background variables that are thought to be related to group membership. Matching method two used propensity score matching where the selection variable and five other variables were included in the logistic regression. Matching method three used propensity score matching but the selection variable was not included in the logistic regression. The third matching method was not expected to work as well as the first two methods because the selection variable was not included in the matching process. However, as described in the next two sections, some of the other background variables were modestly related to the selection variable.

Relationship of Background Variables

Six background variables were used in this study: gender, a dichotomous fee indicator representing low income status, US region, high school grade level, ethnicity, and parental education, which was the selection variable. Table 4.36 provides phi coefficients for each pair of background variables. Phi was calculated by SAS for variables with two or more categories as the square root of the Pearson chi-square divided by the total number of observations (Conover, 1998, p. 234). Phi was calculated using the full set of examinees from each exam that had valid codes for all six background variables. The full set of contingency tables was not provided, but the tables were used to describe the coefficients in Table 4.36 in the next several paragraphs.

The gender ratio was essentially the same for fee reduction, US region, grade level, and parental education (phi was approximately zero for all exams). The phi values across exams for gender and ethnicity were slightly higher than zero because a slightly higher proportion of African American women (or a slightly lower proportion of African American men) took the AP Exams compared to the proportion of women (or of men) in other ethnic groups. Small values of phi for fee reduction and US region appear to be caused by a higher relative proportion of examinees that received a fee reduction in the South, and especially in the West, compared to the proportion of examinees that received a fee reduction in the Northeast or Midwest. The proportion of examinees that received a fee reduction was nearly the same across grade levels (phi values were near zero). There was a much larger relationship between fee reduction and parental education, and between fee reduction and ethnicity. The higher the parental education level, the less likely an examinee received a fee reduction. Nearly 33% of examinees in the lowest category of parental education received a fee reduction. Only about 2% of examinees in the highest category received a fee reduction. Mexican American examinees were the most likely to have a fee reduction (35-40% across the four exams), whereas Asian American and white examinees were the least likely to have a fee reduction.

For English Language, the Midwest had a higher proportion of examinees that took the exam in grade 12 and a smaller proportion that took the exam in grade 11 than in the other regions. For French Language, both the Midwest and the Northeast had a higher proportion of examinees that took the exam in grade 12 and a smaller proportion that took the exam in grade 11 compared to the South and West regions. The West had a larger proportion of examinees that took the Physics B Exam in grade 11 and a smaller proportion of examinees that took the Physics B Exam in grade 12 compared to the other regions. Similarly, a greater proportion of examinees took the Chemistry Exam in grades 10 and 11 and fewer examinees took the exam in grade 12 compared to examinees in other regions.

Across all exams, the West had the largest proportion of examinees in the lowest parental education category and the Northeast had the highest proportion of examinees in the highest parental education category.

Ethnic groups were not uniformly distributed across regions. Across all exams, proportionally more white examinees and fewer Mexican American examinees were in the Northeast compared to the other ethnic groups. More African Americans and fewer Asians were in the South; proportionally more white and fewer Hispanic examinees were in the Midwest; and proportionally more Mexican Americans and fewer African Americans were in the West.

For all exams except English Language, there was a slightly higher proportion of examinees in the top two parental education categories that took the exams in 11th grade, and a slightly higher proportion of examinees in the bottom two parental education categories that took the exams in 12th grade. For English Language there did not appear to be a relationship between parental education and grade level.

The relationship between grade level and ethnicity varied by exam. For English Language, the proportion of African American and White examinees that took the exam in 11th grade was slightly lower and the proportion that took the exam in 12th grade was slightly higher than the proportions of examinees for other ethnicities. For French Language, a slightly higher proportion of Hispanic examinees took the exam in 10th grade, a slightly higher proportion of African American and Asian American examinees took the exam in 11th grade, and a slightly higher proportion of Mexican American and white examinees took the exam in 12th grade. For Chemistry, a slightly higher proportion of Mexican and Asian American examinees took the exam in 10th and 11th grades. A lower proportion of Mexican and Asian American examinees took the Chemistry exam in grade 12. A slightly lower proportion of African American examinees took the exam in 11th grade, and a higher proportion of African American examinees took the exam in 12th grade. A greater proportion of Asian American examinees took the Physics B Exam in grades 10 and 11, and therefore a smaller proportion of Asian American examinees took the Physics B Exam in grade 12 compared to other ethnic groups. A higher proportion of Hispanic examinees took the Physics B Exam in grade 12 and therefore a lower proportion of Hispanic examinees took the Physics B Exam in grade 11.

Across all exams, there was a moderate relationship between parental education and ethnicity. White and Asian American ethnic groups had the highest proportion of examinees in the highest parental education categories and the lowest proportion of examinees in the lowest parental education categories. The Mexican American ethnic group had the highest proportion of examinees in the lowest parental education categories and the lowest proportion of examinees in the highest parental education categories.

Logistic Regression Results

Matched sampling was conducted for matching methods M_2 and M_3 by matching the new group to the old group in terms of propensity scores, or the probability an examinee was part of the old form group given their set of scores on the background variables entered into a logistic regression equation. For M_2, six background variables were entered into the logistic regression equation: gender, ethnicity, grade level, fee indicator, US region, and parental education. For M_3, parental education was not included in the logistic regression equation. Because the unmatched groups were sampled to differ in terms of the average level of parental education, it was expected that the regression coefficient for parental education would be significant. Coefficients for the other variables were not expected to be significant because the variables should have been represented in approximately randomly equivalent proportions in the old and new form groups. However, as indicated in Table 4.36, the six background variables were not unrelated. In fact, the phi coefficients for fee and parental education were moderate (0.30-0.35), as were the coefficients for ethnicity and parental education (0.28-0.44).

Tables 4.37-4.40 provide the generalized (pseudo) r-squared values (Cox & Snell, 1989, p. 208-209) and p-values for the logistic regression coefficients for the intercept and the background variables entered into the model. The same patterns were found across all exams. As the ES increased, the r-squared values increased, indicating that background variables predicted group membership better for higher ESs. For M_2, this finding was expected because group differences in terms of parental education increased as ES increased. For M_3, the r-squared values were substantially lower than those for M_2, which was expected because the selection variable was not included in the model. However, the r-squared values increased as ES increased for M_3 as well, which may have been caused by the correlation between parental education and the other background variables. The r-squared values were fairly low for M_2, even at an ES of 0.3. The highest r-squared value at ES=0.3 was 0.2678 for Physics B (see Table 4.40). The r-squared value at ES=0.75 for English Language was 0.6260 (see Table 4.38).

For M_2, the intercept and parental education regression coefficients were always significantly different from zero ($p<0.05$). The coefficient for the reduced exam fee indicator was also significant in a majority of cases. For M_3, findings were somewhat mixed. For Chemistry, the coefficient for fee was significant for ESs of 0.2 and 0.3. For English Language, coefficients for all variables except high school grade level were significant. For French Language, coefficients for grade, fee, and the intercept were often significant. For Physics B, the coefficients for fee and grade were significant in most cases.

Matching Results

Of the three matching methods, methods M_1 and M_2 were expected to provide more similar results because both included the selection variable, parental education. But how did the results compare to those obtained with M_0 and M_3? In this section, the number of examinees at each level of parental education is compared for the four sampling methods. The ES differences between old and new form groups for each of the methods are also compared. In later sections, the results are compared in terms of the accuracy of equating results (Research Question 4) and the impact on violations of equating assumptions (Research Question 5).

Frequencies across Levels of Parental Education

Tables 4.41-4.45 provide the number of examinees within each level of parental education based on matching methods M_0-M_3. Recall that M_0 is the unmatched condition. In general, for M_0, the Form A frequencies within the higher levels of parental education stayed the same or decreased as ES increased. Also, typically for M_0, the Form B frequencies within the lower levels of parental education stayed the same or decreased as ES increased. For M_1, frequencies were the same for Forms A and B for each level of parental education. Comparing M_2 to M_0 and M_1, it appears that M_2 resulted in more similar frequencies across the levels of parental education for Forms A and B. In fact, for English Language, M_2 provided the same frequencies as $M_1$for Forms A and B. It appears that as ES increased M_2 provided a closer approximation to M_1 in terms of frequencies although for Physics B, the lowest ES had perfect agreement between M_1 and M_2 (see Table 4.44).

The extremely low frequencies for English Language ES=0.5, especially the M_0 and M_3 frequencies were a result of an attempt to find a combination of frequencies across the four levels of parental education that would result in such an extreme ES. However, a different ratio of frequencies, with larger frequencies at the low parental education levels, would have been possible.

A comparison of M_3 with M_0 and M_1 indicates that for both forms, M_3 frequencies were much closer to M_0 frequencies than to M_1 frequencies. For Form B, M_3

and M_0 were nearly identical for all ESs and exams. For Form A, M_3 was typically less than M_0 for parental education categories 1 and 2, and greater than M_0 for parental education categories 3 and 4. These findings were expected given that M_3 matched the new form group to the old form group using propensity scores that were not highly related to group membership (recall the small pseudo r-squared values for M_3 in Tables 4.37-4.40). M_3 made new and old form groups slightly more similar in terms of the number of examinees within each level of parental education, but was much less effective than the M_1 and M_2 matching methods.

Effect Sizes based on Four Matching Methods

Table 4.45 provides the target common item ES for M_0 and the observed common item ES for the M_0, M_1, M_2, and M_3 matching methods. In general, the observed ESs for M_0 were close to the target ESs. The observed ESs for English Language and Physics B were all slightly smaller than the target ESs for M_0. The ESs for the M_1 and M_2 matching methods were all relatively small, indicating that the groups were much more similar after matching than before matching. This finding was expected given the similarity of frequencies across levels of parental education for M_1 and M_2. The M_3 ESs were very similar to the M_0 ESs. In many, but not all cases, the M_3 method did reduce the ES slightly compared to the original unmatched ES. However, the M_3 method was not successful in making the groups nearly identical in terms of common item performance. This finding was expected given the similarity of frequencies across levels of parental education for M_0 and M_3.

Research Question 4

Which matching techniques, if any, provide more accurate equating results?

In Research Question 2, it was found that the equating results for the four equating methods became less similar as ES increased (although the two IRT methods tended to be very similar regardless of ES). However, within a given method, the accuracy of comparison equating results, in terms of the closeness of the results to the criterion equating results, was not always related to the magnitude of the ES difference at least for ESs of 0.3 or less. If the ES does not directly impact the accuracy of equating, then matching methods which, in effect, decrease the ES, may not improve equating accuracy. In this section, equating results are compared using the same four methods used in Research Question 2: comparisons of equating relationships, old form equivalent moments, REMSD, and classification consistency.

Comparison of Equating Relationships

In Research Question 2, the impact of group differences on equating results was investigated for four AP Exams using common item ES differences ranging from 0 to 0.75. A comparison of difference plots indicated that differences in the equating results for the traditional equating methods increased as ESs increased, but very little difference in IRT equating results occurred, regardless of the magnitude of the ES. However, the hypothesis that increases in ES would result in increases in the divergence of comparison and criterion equating results for a given equating method was only substantiated with very extreme ESs, as were used for English Language. With ESs ranging only from 0 to 0.3, there was no consistent pattern as between the magnitude of the ES and the divergence of equating results from the criterion (ES=0).

For Research Question 4, matched and unmatched results were compared. For each exam and equating method, results from 12 ES-matching combinations were compared. Difference plots comparing the old form equivalents for each matching method compared to the criterion equating relationship (ES=0), are provided for Chemistry in Figures 4.38 and 4.39. In Figure 4.38, difference plots are provided for the IRT true score equating method in the left column and for the IRT observed score equating method in the right column. The top difference plots are for an ES of 0.1, the middle for an ES of 0.2, and the bottom for an ES of 0.3. The equating relationships are plotted for approximately the first through the 99th percentiles. Equating relationships are provided in each plot for the unmatched condition (M_0), and the three matched conditions (M_1-M_3) using colored lines. The closer the equating relationships are to the vertical axis value of zero, the closer they are to the criterion equating relationship where ES=0. The two black lines represent plus and minus two SEs as calculated using 1000 bootstrap replications for the criterion chained equipercentile equating. Although these SEs may not be the same as the IRT bootstrap SEs, they provide some indication of how different the equating relationships are. The same information is provided for the traditional equating methods for Chemistry in Figure 4.39. On the left are the results for frequency estimation, and on the right are the results for chained equipercentile.

A comparison of the difference plots in Figure 4.38 shows that for Chemistry, the IRT methods produced nearly identical results across all ESs. M_0 appeared to deviate most from the criterion for ES=0.1 but not for the other ES levels. All matching methods for ES= 0.2 and 0.3 provided similar results mostly within plus or minus two SEs, except at the low end of the scale. A comparison of Figures 4.38 and 4.39 indicates that the traditional methods did not provide results that were as similar as the IRT method results. M_0 appeared most deviant at ES=0.1 but it was unclear using graphical inspection which matching method was the most deviant at other ES levels. All matching methods provided traditional equating results that were within plus or minus two SEs for the majority of the score range. However, the variability in the equating results produced by the four matching methods was much greater for the traditional methods than for the IRT methods. The largest difference between equating relationships appeared to be approximately six score points or a difference of approximately 0.13 standard deviations.

Figures 4.40 and 4.41 provide difference plots for English Language IRT and traditional equating methods, respectively. The rows of Figures 4.40 and 4.41 provide difference plots for ESs of 0.25, 0.50, and 0.75. Note that the vertical scale was -20 to 20 in order to accommodate larger equating differences. (For Chemistry the vertical scale was -10 to 10.) Comparing the left three plots in Figure 4.40 to the right three plots indicates that both IRT methods provided nearly identical results. As ES increased, the deviation of M_0 and M_3 equating relationships from the criterion (vertical axis value of zero) increased. At ES=0.25, the M_0 and M_3 equating relationships differed from the criterion by nearly 10 score points at the low end of the scale; by ES=0.75, they differed by nearly 20 score points. For ES=0.25 and 0.50, M_1 produced the closest equating relationship to the criterion, and results that were within plus or minus two SEs from the criterion equating relationship for the majority of the score scale, even at ES=0.75. M_2 produced equating results that were very close to the criterion for ES=0.75, but comparable to M_0 and M_3 for ES= 0.25 and 0.50. In fact, M_2 deviated the most from the criterion for ES= 0.25 and 0.50 at the high end of the scale. The sensitivity of the IRT equating results to the matching method appeared to increase as ES increased. By ES=0.75, the equating relationship for M_2 and M_0 differed by almost 15 points, or a half of a standard deviation, at the low end of the scale.

A comparison of the frequency estimation and chained equipercentile equating results in Figure 4.41 indicates that the equating results for the traditional methods were not nearly as similar as the equating results for the two IRT equating methods. At ES=0.25, M_0 and M_3 deviated from the criterion by more than two SEs for the majority of the score range for the frequency estimation method, and at some score points for the chained equipercentile method. The deviation of M_0 and M_3 from the criterion increased as the ES increases for both methods, but the frequency estimation method appeared more sensitive to the matching method. By ES=0.75, M_0 differed from the criterion by approximately 15 points (approximately 0.5 SD) at some scores for the frequency estimation method. M_1 and M_2 tended to stay within plus or minus two SEs even at ES=0.75.

Difference plots for the IRT methods are provided in Figure 4.42 for French Language. M_0 tended to deviate the most for ES=0.1 and 0.3, especially at the low end of the scale. M_1 differed the most from the criterion at low scores for ES=0.2. All matching methods provided results that were within plus or minus two SEs of the criterion equating relationship, even at ES=0.3. Again, both IRT true score and observed score equating methods provided nearly identical results.

Traditional equating difference plots are provided in Figure 4.43 for French Language. At ES=0.1, the traditional equating results for all matching methods were within plus or minus two SEs for a majority of the score scale. For frequency estimation, the M_0 and M_1 methods deviated from the criterion by more than two SEs at several score points for ES=0.2. At ES=0.3, the M_0 and M_3 methods deviated from the criterion by more than two SEs in the middle range of scores. For chained equipercentile, even at ES=0.3, all matching methods provided equating results within plus or minus two SEs of the criterion equating relationship for the majority of the score scale. The equating results for the traditional methods were much less similar than the equating results for the IRT methods. The largest difference in the equating results produced by the four matching methods was approximately 5 score points or 0.14 standard deviations.

Finally, Figures 4.44 and 4.45 provide the Physics B IRT and traditional equating results respectively. Because the composite scale was so much shorter for Physics B, the standard deviation was much smaller than the standard deviation for the other exams. The differences in equating relationships were similarly smaller, and so the difference plots have a vertical scale ranging from -5 to 5 instead of -10 to 10 as was used with Chemistry and French Language.

Comparison of the IRT true score and observed score results (Figure 4.44), indicates that the two methods provided nearly identical results. The accuracy of the matching methods in terms of the criterion equating relationship was not consistent across ESs. At ES=0.1, only M_3 was within two SEs of the criterion equating relationship for the entire score range. At ES=0.2, M_1 and M_2 deviated by more than two SEs, and at ES=0.3 all matching methods were within plus or minus two SEs of the criterion. For the traditional methods (see Figure 4.45), M_1 and M_2 exceeded two SEs at the high end of the scale for all comparison ESs. At ES=0.3, M_0 exceeded two SEs for a large number of score points for the frequency estimation method, and a few score points for the chained equipercentile method. The largest difference between matching methods was approximately 3 score points, or 0.16 standard deviations.

Moments

As part of Research Question 2, it was noted that as ES increased, the frequency estimation old form equivalent means became increasingly higher than the IRT true and observed score equated means. The means for the chained method tended to fall in between the frequency estimation means and the IRT method means. Both IRT methods provided similar moments even at an ES of 0.75. For Research Question 4, the unmatched and matched equating results are compared to see whether or not matching results in more similar moments for the four equating methods as ES increases.

Moments are provided for each combination of ES, matching method, equating method, and exam, in Tables 4.46-4.49. Because smoothed and unsmoothed moments were very similar, only the smoothed moments were provided for the frequency estimation and chained equipercentile equating methods. Also, because the old form group was sampled to be higher performing than the new form group for all but the ES=0 condition, there was no criterion mean. That is, the moments found for the comparison conditions were not expected to be the same as the criterion mean, even when the ES for the matched groups was near zero. Therefore the ES=0 moments are not provided in Tables 4.46-4.49 because they were provided in Tables 4.14-4.17.

Though there was no “true mean” that could be used to determine the accuracy of the equated means in Tables 4.46-4.49, predictions can be made about the relationship between the old form means and the old form equivalent means for each condition. It was already noted in the result section for Research Question 2 that when the ES was zero, all methods produced similar results. When the ES increased and groups were not matched (M_0), there were two general findings: 1) because the old form group was sampled to be higher performing than the new form group, as ES increased, old form equivalent means decreased, and 2) as ES increased, frequency estimation and chained equipercentile means became increasingly higher than IRT means, although the chained equipercentile means were lower than frequency estimation means.

For Research Question 4, old form equivalent means for different ESs and equating methods were compared for each matching method. For M_1, if matching was effective, old form means and old form equivalent means should be similar because the M_1 matching process decreases the old form group average, and increases the new form group average. Therefore, the M_1 old form means should be lower than the M_0 old form means. Recall that for M_1, the number of examinees sampled from each parental education level was the smaller of the two sample sizes in the old and new form groups. If, for example, there were 500 examinees in parental education category one for the old form group, and 700 examinees in parental education category one for the new form group, then all 500 examinees in the old form group would be retained in the matched sample, and 500 examinees would be randomly sampled from the first parental education category for the new form group matched sample. There were always more examinees in the lower parental education categories for the unmatched new form group, and more examinees in the higher parental education categories for the unmatched old form group. However, when the groups were matched using M_1, there were the same number of examinees in each parental education category. The difference between the groups in terms of performance should be nearly zero. Therefore, the only difference between the old form mean and the old form equivalent means for M_1 would be caused by small remaining group differences.

For M_2, the new group was matched to the old group based on propensity scores. If there was a match for each old form group examinee, then the old form mean would be the same as in M_0 and the old form equivalent means would be similar to the old form mean. However, the examinees in the old form group that did not have matches were likely to be the highest performing because the new form group was lower performing. Therefore, the old form group mean is likely to be lower than the M_0 old form mean and the old form equivalent means will be similar to the old form mean to the extent to which the matching resulted in equivalent groups.

For M_3, the same logic applies as for M_2 except that the propensity scores did not include parental education. If the M_3 matching method does not result in more similar groups in terms of exam performance, then the M_3 old form mean is likely to be similar to the M_0 old form mean, and the M_3 old form equivalent mean is likely to be similar to the M_0 old form equivalent mean.

A comparison of old form means and old form equivalent means in Tables 4.46-4.49 shows the expected patterns just described. For example, for Chemistry (Table 4.46), the M_0 old form mean for ES=0.1 was approximately 108 and the old form equivalent means were all around 105. The higher old form mean was expected because the old form group was approximately 0.1 SDs higher than the new form group in terms of common item performance. The M_1 old form mean and old form equivalent means were all approximately 105 for Chemistry, indicating that the M_1 matching method resulted in very similar groups. The M_2 old form means and old form equivalents were less similar, and as expected, because the M_3 method did not provide a good match for the old and new form groups, the old form mean was very similar to the M_0 old form mean, and the M_3 old form equivalent means were similar to the M_0 old form equivalent means. This pattern held for ES=0.2 and 0.3. In general the M_0 old form means increased as ES increased, and the M_0 old form equivalent means decreased as ES increased because the old and new form group performance differences increased. These patterns were consistent across all four exams.

For English Language, where the group differences were more extreme, as ES increased, it appears that the old form equivalent means for the four equating methods became less similar for M_0 and M_3. However, methods M_1 and M_2 resulted in similar old form equivalent means across equating methods even for ES=0.75.

REMSD

M_0 results were compared to REMSD values for M_1-M_3. Tables 4.50-4.53 provide the REMSD values for all ESs, matching methods, equating methods, and exams. Note that the frequency estimation and chained equipercentile REMSD values were based on postsmoothed equivalents. There were two hypotheses about the magnitude of the REMSD values. First, it was expected that REMSD would increase for the unmatched condition (M_0) as ES increased. As noted in Research Question 2, this trend was only found for English Language where the ESs were much more extreme. For the other three exams, there was no obvious trend between ES and REMSD.

The second hypothesis was that the REMSD values would be smaller for M_1 and M_2 than for M_0 and M_3, because the matched samples equating results were expected to be closer to the criterion equating relationships. A comparison of the REMSD values for the four matching methods across all levels of ES, equating method, and exam, indicated that the M_1 and M_2 REMSD values were often (but not always) smaller for the frequency estimation method, and were always smaller for English Language when the ES was 0.5 or 0.75.

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

The exam with the highest REMSD values was English Language, which was expected because of the larger ESs. The average REMSD values for the other exams were fairly similar. Only Chemistry and Physics B had REMSD values below the SDTM. For Chemistry, only the M_1 and M_3 IRT values at ES=0.1 were lower than the SDTM. For Physics B, the frequency estimation REMSD values were below the SDTM for M_3 at ES=0.1, and M_0 at ES=0.2. For the chained equipercentile method, the M_0 REMSD value was lower than the SDTM at ES=0.2. For both IRT true score and observed score equating methods, the M_0 REMSD values at ES=0.3 were both below the SDTM. These results indicate that the matching methods were not successful in lowering the REMSD values below the SDTM threshold.

Classification Consistency

Classification consistency was considered for M_0 as part of Research Question 2. When AP grades based on the criterion equating relationship were compared to AP grades for comparison equating relationships, the magnitude of classification consistency values did not appear to correspond in a predictable way to the magnitude of the ES. Classification consistency for the IRT equating methods was often higher than the classification consistency for the traditional equating methods. Physics B had the highest average classification consistency and English Language had the lowest. When the AP grades based on IRT true score equating were compared to the AP grades based on equating results from the other three equating methods, IRT observed score equating had higher classification consistency than the traditional methods, and classification consistency tended to decrease as ES increased.

For Research Question 4, classification consistency was compared for all four matching conditions. Cut scores are provided for each combination of ES, matching method, equating method, and exam in Tables 4.54-4.69. For example, postsmoothed frequency estimation cut scores for Chemistry are provided in Table 4.54. A comparison of the cut scores for ES=0.1, 0.2, and 0.3 to cut scores for ES=0 provides some indication of how high the classification consistency is between the criterion and comparison equating relationships. In general, it appears that the traditional equating methods were more variable than the IRT methods for all exams except Physics B. As ES increased, the variability of comparison equating cut scores did not appear to change in a predictable way. The most variable matching method was not consistent across exams. For English Language, M_1 and M_2 appeared to provide an improvement over M_0 at ES=0.5 and 0.75. M_3 provided a slight improvement at ES=0.5 and 0.75. Cut scores appeared most variable for English Language, which was expected due to the more extreme ESs. Cut scores appeared least variable for Physics B. However, there were only approximately half as many composite score points for Physics B compared to the other exams.

Tables 4.70-4.73 provide the classification consistency values for each comparison equating where the cuts for the ES=0 equating relationship were considered the criterion. A comparison of classification consistency values for Chemistry (Table 4.70) indicates that M_1- M_3 provided an improvement over M_0 at ES=0.1 but not consistently at ES=0.2 and 0.3. There did not appear to be a consistent relationship between ES and classification consistency. Across all conditions, classification consistency ranged from 93.90 to 100%.

Table 4.71 provides the classification consistency values for English Language. The highest classification consistency values were for ES=0.25. However, classification consistency did not decrease consistently for ES=0.5 and 0.75. IRT equating methods tended to have the highest classification consistency values. M_0 and M_3 tended to have the lowest classification consistency values, especially for the frequency estimation method. M_1 and M_2 tended to provide an improvement over M_0, most noticeably for the frequency estimation method. M_3 did not provide an improvement in classification consistency values over M_0. Across all conditions, classification consistency ranged from 70.11-98.82%

Classification consistency values for French Language are provided in Table 4.72. In general, the IRT methods tended to have higher classification consistency values. M_0 and M_3 tended to have the lowest classification consistency. M_1 and M_2 typically provided an improvement over M_0. On average, classification consistency tended to decrease as ES increased, although for any given equating method and matching method, the trend was not completely consistent. Across all conditions, classification consistency ranged from 93.38-100%.

Classification consistency values for Physics B are provided in Table 4.73. The patterns for Physics B were not the same as those seen with the other exams. The frequency estimation method provided the highest classification consistency in a majority of cases, instead of the IRT methods. M_1 and M_2 did not provide an improvement over M_0. Finally, classification consistency did not appear to decrease as ES increased. Classification consistency across all conditions for Physics B ranged from 91.29-100%.

In Tables 4.74-4.77, classification consistency values are provided for the four exams with IRT true score equating as the criterion equating relationship. These classification consistency values indicate the similarity of the equating results across methods, not the accuracy of equating results. Across all four exams, the classification consistency values for IRT observed score equating were very high (range=98.41-100%), regardless of the matching method. The results for the traditional equating methods were not nearly as similar (range=75.39-100%). For Chemistry, French Language, and Physics B (Tables 4.74, 4.76, and 4.78), the classification consistency values for M_0 tended to decrease as ES increased, but the classification consistency values for matching methods M_1-M_3 were not consistent across ES, nor was it clear that matching methods M_1-M_3 provided more consistent results than M_0. However, for English Language (Table 4.75), M_1 and M_2 tended to result in more consistent classifications for the traditional methods. M_0 and M_3 provided less consistent results, especially for the frequency estimation method.

In general, the matching method appeared to only affect classification consistency in a predictable way when the ES was large. When the ES was large, matching methods M_1 and M_2 provided more consistency of AP grades across equating methods, and more consistency of AP grades when compared to the AP grades examinees received using the criterion equating relationship (ES=0).

Research Question 5

Can matched samples equating reduce the extent to which equating assumptions are violated?

Frequency Estimation

As described for Research Question 3, a weighted absolute maximum difference between the old form group and new form group cumulative frequency distributions was used to evaluate the frequency estimation assumptions. The weighted absolute maximum differences are provided in Tables 4.78-4.81 for the four exams. The third and fourth columns of the four tables provide nearly the same numbers, indicating that the degree to which the old and new form groups differed in terms of their conditional distributions on Form A was similar to the degree to which the groups differed in terms of their conditional distributions on Form B. As ES increased, the weighted maximum differences increased for both forms for M_0. In other words, as groups differences increased, the adequacy of frequency estimation assumptions decreased. However, the weighted maximum differences did not increase for M_1 and M_2 as ES increased. In fact, the weighted maximum differences were almost always smaller for M_1 and M_2, than for M_0. Weighted maximum differences for M_3 were usually larger than for M_0, although the M_3 matching method did appear to improve the degree to which the frequency estimation assumptions held for English Language. However, the M_1 and M_2 matching methods provided better results than M_3 for all exams.

The values in Tables 4.78-4.81 were compared to the REMSD values for the frequency estimation method in Tables 4.50-4.53. It was expected that higher weighted maximum differences (assumptions holding less well) in Tables 4.78-4.81 would lead to higher REMSD values (less accurate equating results) for the frequency estimation method. However, REMSD values and weighted maximum differences only showed the expected pattern for English Language and French Language. For Chemistry and Physics B, the degree to which the frequency estimation assumptions held did not appear to correspond to equating accuracy as measured by the REMSD statistics.

Chained Equipercentile

As described in Research Question 3, chained equipercentile equating assumptions were assessed by calculating REMSD values comparing the linking relationship between the composite and common item scores for old and new form groups. REMSD statistics are provided in Tables 4.82-4.85. The REMSD values in columns 3 and 4 were fairly similar indicating that the impact of different ESs and matching methods on the similarity of *A* to *V* equating results for old and new form groups was comparable to the impact of different ESs and matching methods on the similarity of *V* to *B* equating results for the two groups.

A comparison of the REMSD values for M_0 indicates that they increased consistently as ES increased for English Language, but for the other exams, the ES and REMSD values did not appear related. For Chemistry (Table 4.82), M_0 had the biggest REMSD values in most cases M_1-M_3 all had smaller REMSD values, indicating that the chained equipercentile assumptions appeared to hold better in the matched samples, than in the unmatched samples. For English Language (Table 4.83), REMSD values for M_1 were smallest across all ESs. REMSD values for M_2 only provided an improvement over

M_0 values at ES=0.5 and 0.75. The M_3 matching method did not appear to improve results over the M_0 method. For French Language (Table 4.84), no matching method had consistently lower REMSD values across all ES levels. For Physics B (Table 4.85), M_1-M_3 had smaller REMSD values compared to M_0 for ES=0.1 and 0.3, but not for 0.2.

As found in Research Question 3, the REMSD values in Tables 4.82-4.85 corresponded closely to the chained REMSD values in Tables 4.50-4.53, for M_0. The values also corresponded closely for the other matching methods for all exams except Physics B. The similarity of the REMSD values indicates that the degree to which the chained assumptions held was directly related to the closeness of the comparison and criterion equating relationships. This finding highlights the fact that equating results are more accurate when equating method assumptions hold.

IRT

Correlational Analyses

In Research Question 3, three types of correlational analyses were used to evaluate the IRT assumption of unidimensionality: MC-FR correlations, principal components analysis, and PolyDIMTEST/PolyDETECT software. In Table 4.27 the MC and FR section reliabilities and observed and disattenuated correlations were provided for the M_0 matching method. ES did not have any predictable influence on the magnitude of the correlations. A comparison of the observed and disattenuated correlations for M_1-M_3 likewise did not reveal any systematic differences. Minimum and maximum observed and disattenuated correlations are provided for all four exams in Table 4.86. As noted for Research Question 3, the Chemistry and Physics B Exams appeared to be most unidimensional, in that their MC-FR correlations were closest to 1. The maximum disattenuated correlation found for English Language was the lowest at 0.755, which indicates that the English Language Exam may not be unidimensional.

In Tables 4.28-4.31 the first five eigenvalues and percentage of variance explained were provided for the M_0 matching condition as part of Research Question 3. Scree plots for the ES=0 condition were provided in Figures 4.28-4.31. It was noted that the eigenvalues and scree plots were nearly identical across ESs for M_0. The eigenvalues and scree plots for M_1-M_3 were likewise nearly identical to M_0 values. Therefore, eigenvalues and scree plots were not provided for M_1-M_3. There did not appear to be any relationship between the dimensionality of the exams, as assessed by principal components analysis, and the matching method used. All four exams appeared to have one dominant dimension and one or two minor dimensions based on evaluation of the eigenvalues and scree plots.

Because of the limitations of both PolyDIMTEST and PolyDETECT, the programs were not used with Forms A and B of the four AP Exams.

Item Parameter Estimates

As part of Research Question 3, the MC and FR item parameters were estimated together and separately to assess the stability of the results. Originally, it was predicted that a more homogeneous set of items (only MC or only FR items) would have higher discrimination estimates than when the MC and FR item parameters were estimated together. In fact, it was found that average MC item *a*-, *b*-, and *c*-parameter estimates increased when estimated with the FR items. For FR items, the average discrimination decreased when estimated with the MC items, but the average *b*-parameter estimates did not follow a consistent pattern across forms or exams. Another finding was that the average *b*-parameter estimates increased as ES increased for Form A, and the average *b*-parameter estimates decreased as ES increased for Form B (see Tables 4.32-4.35).

For Research Question 5, the MC and FR item parameter estimates were estimated together and separately, and the process was repeated for each of the matching conditions (M_0-M_3). Additional scatter plots of item parameter estimates were not included for Research Question 5 because they were very similar to those included for Research Question 3 (see Figures 4.22-4.37).For Chemistry, average MC parameter estimates are provided in Table 4.87 and average FR parameter estimates are provided in Table 4.88. Similar tables are provided for the other three exams in Tables 4.89-4.94. A comparison of the difference (Diff) columns indicates that most average MC item parameters increased when estimated with the FR items for all exams (i.e., the differences were all positive). However, for French Language Form A, the average MC *a*-parameter estimates often decreased when estimated with the FR items (see Table 4.91). Also, a few of the average MC *a*-parameter estimates decreased when estimated with the FR items for English Language (see Table 4.89).

As noted in the results section for Research Question 3, the average discrimination values for the FR section decreased when estimated with MC items. This pattern was found consistently across matching methods M_0-M_3. However, the finding that average *b*-parameter estimates for both FR and MC items increased as ES increased for Form A was not always found when matching methods M_1-M_3 were used. For example, the average Chemistry M_0 MC *b*-parameter estimates, estimated simultaneously with FR items (see Table 4.88, column 6), increased from -0.0068 to 0.083 for Form A. However, for M_1, the values for ESs of 0.1, 0.2, and 0.3 were -0.0542, 0.0056, and -0.0886 respectively, indicating no obvious increasing or decreasing trend. Though the matching methods appeared to change the pattern of average MC and FR *b*-parameter estimates, the parameter estimates still fluctuated noticeably across ES levels.

Comparison of Equating Results

Equating methods used with the CINEG design have much more stringent statistical assumptions than equating methods used with the SG design, especially for the traditional equating methods. Differences between CINEG and SG equating results provides indirect evidence that the equating assumptions involved with the CINEG equating methods may not hold. In the first section of the results chapter, the CINEG and SG results were compared for SG data. The results were nearly identical. Therefore, the comparison of equating results for comparison equatings (M_0-M_3) to the criterion equating relationship (ES=0) provides an indication of how well the CINEG equating assumptions hold for different types of matched samples. These comparisons were made above, to answer Research Question 4. For M_0, there was indirect evidence that the assumptions may have been violated for all of the equating methods because as ES increased, there was a tendency for the equating results to become further away from the criterion equating results. This finding was most apparent for English Language which had the most extreme ESs. However, at the most extreme ESs of English Language, matching methods M_1-M_3 appeared to make the comparison equating results closer to the criterion equating results. Matching methods M_1 and M_2 provided more improvement than M_3, indicating that matching methods that include the selection variable may decrease the degree to which equating assumptions are violated.

Table 4.1 Comparison of Old Form Equivalent Composite Score Moments using SG and CINEG Traditional Equating Methods

		SG		CINEG			
Exam	Moments	UEq	SEq	UFE	SFE	UCE	SCE
Chem	Mean	106.05723	106.06663	106.05555	106.06377	106.05488	106.06872
	SD	44.5235	44.51783	44.54002	44.50978	44.54141	44.49479
	Skew	-0.27775	-0.27729	-0.27782	-0.27743	-0.27794	-0.27749
	Kurt	1.99436	1.99279	1.99725	1.99282	1.99752	1.99318
English	Mean	125.81016	125.80605	125.81027	125.81103	125.81027	125.79712
	SD	29.47225	29.48398	29.44888	29.47568	29.44888	29.52839
	Skew	-0.4622	-0.46324	-0.46393	-0.46367	-0.46393	-0.48555
	Kurt	3.26917	3.27933	3.25857	3.27406	3.25857	3.33886
French	Mean	128.41334	128.41569	128.43223	128.41932	128.43023	128.41283
	SD	35.62689	35.61876	35.57346	35.60564	35.56875	35.60869
	Skew	-0.30955	-0.30979	-0.30024	-0.30899	-0.30113	-0.30989
	Kurt	2.69404	2.69269	2.6669	2.68699	2.6653	2.67935
Physics	Mean	42.23355	42.23174	42.23082	42.23142	42.23082	42.22815
	SD	18.60597	18.60003	18.60434	18.59655	18.60434	18.59268
	Skew	0.07469	0.07208	0.07322	0.07247	0.07322	0.07229
	Kurt	2.12536	2.11943	2.12364	2.11987	2.12364	2.11771

Note. UEq = Unsmoothed Equipercentile, SEq = Smoothed Equipercentile, UFE= Unsmoothed Frequency Estimation, SFE=Smoothed Frequency Estimation, UCE=Unsmoothed Chained Equipercentile, SCE=Smoothed Chained Equipercentile.

Table 4.2 Comparison of Old Form Equivalent Composite Score Moments using SG and CINEG IRT Equating Methods

		IRT True		IRT Observed	
Exam	Moments	SG	CINEG	SG	CINEG
Chem	Mean	106.27535	106.27986	106.30606	106.31768
	SD	44.69925	44.89784	44.68283	44.87044
	Skew	-0.29268	-0.30355	-0.28872	-0.29881
	Kurt	1.99917	2.00172	2.00104	2.00265
English	Mean	125.83168	125.75657	125.80839	125.75300
	SD	29.63048	29.67692	29.59894	29.62035
	Skew	-0.44237	-0.47427	-0.45273	-0.47743
	Kurt	3.17511	3.25117	3.21366	3.27361
French	Mean	128.40624	128.65959	128.41686	128.63843
	SD	35.61225	35.05571	35.63236	35.13043
	Skew	-0.28886	-0.19572	-0.28850	-0.20322
	Kurt	2.65896	2.57690	2.65083	2.57853
Physics	Mean	42.31877	42.37933	42.33253	42.39278
	SD	18.70782	18.72282	18.74043	18.75413
	Skew	0.06169	0.06933	0.06151	0.06942
	Kurt	2.11758	2.11893	2.11757	2.11889

Table 4.3 Chemistry Old Form Equivalent Composite Score Moments for Parental Education Subgroups

Parent ED Level	Moments	UEq	SEq	IRT True	IRT Obs
Total	Mean	106.05723	106.06663	106.27535	106.30606
	SD	44.52350	44.51783	44.69925	44.68283
	Skew	-0.27775	-0.27729	-0.29268	-0.28872
	Kurt	1.99436	1.99279	1.99917	2.00104
1 & 2	Mean	82.69912	82.69259	82.83222	82.81508
	SD	44.73495	44.72241	44.60104	44.59937
	Skew	0.22694	0.22596	0.18541	0.18914
	Kurt	1.88501	1.88268	1.87251	1.87910
3	Mean	105.46458	105.46768	105.60880	105.65220
	SD	41.81212	41.80402	41.85358	41.80591
	Skew	-0.24407	-0.24330	-0.27610	-0.26675
	Kurt	2.12654	2.12514	2.15749	2.15444
4	Mean	121.89215	121.9036	121.97441	122.04746
	SD	39.93514	39.93407	40.17914	40.13008
	Skew	-0.70649	-0.70710	-0.68985	-0.68077
	Kurt	2.73695	2.73768	2.64893	2.63624

Table 4.4 English Language Composite Score Moments for Parental Education Subgroups

Parent ED Level	Moments	UEq	SEq	IRT True	IRT Obs
Total	Mean	125.81016	125.80605	125.83168	125.80839
	SD	29.47225	29.48398	29.63048	29.59894
	Skew	-0.46220	-0.46324	-0.44237	-0.45273
	Kurt	3.26917	3.27933	3.17511	3.21366
1	Mean	104.6292	104.62917	104.79972	104.69463
	SD	31.37947	31.36801	31.40793	31.47824
	Skew	-0.26842	-0.26773	-0.24350	-0.26869
	Kurt	2.76121	2.75763	2.78381	2.82340
2	Mean	114.99236	114.99027	115.13801	115.05755
	SD	28.75666	28.75208	28.72614	28.75391
	Skew	-0.36997	-0.37277	-0.36468	-0.38966
	Kurt	2.96265	2.95915	2.95069	3.00683
3	Mean	127.35628	127.36443	127.39861	127.37390
	SD	28.06151	28.03412	28.14101	28.11854
	Skew	-0.56246	-0.55429	-0.51666	-0.52985
	Kurt	3.72542	3.68673	3.54033	3.58814
4	Mean	135.50719	135.51836	135.48742	135.48456
	SD	28.28593	28.24432	28.45478	28.39719
	Skew	-0.57090	-0.56112	-0.54009	-0.54675
	Kurt	3.39561	3.34920	3.29427	3.32288

Table 4.5 French Language Old Form Equivalent Composite Score Moments for Parental Education Subgroups

Parent ED Level	Moments	UEq	SEq	IRT True	IRT Obs
Total	Mean	128.41334	128.41569	128.40624	128.41686
	SD	35.62689	35.61876	35.61225	35.63236
	Skew	-0.30955	-0.30979	-0.28886	-0.28850
	Kurt	2.69404	2.69269	2.65896	2.65083
1 & 2	Mean	108.27525	108.27694	108.30582	108.31130
	SD	36.23122	36.24746	36.36743	36.40312
	Skew	0.03702	0.03701	0.04945	0.04722
	Kurt	2.41697	2.42082	2.39299	2.39367
3	Mean	124.77937	124.78149	124.79172	124.80717
	SD	34.05609	34.05529	33.96288	33.99484
	Skew	-0.22965	-0.22786	-0.20258	-0.20284
	Kurt	2.68071	2.67920	2.71003	2.70495
4	Mean	137.51442	137.51597	137.47668	137.47461
	SD	33.28558	33.26598	33.11317	33.15522
	Skew	-0.40126	-0.40022	-0.36100	-0.36122
	Kurt	2.96192	2.95178	2.93062	2.92396

Table 4.6 Physics B Old Form Equivalent Composite Score Moments for Parental Education Subgroups

Parent ED Level	Moments	UEq	SEq	IRT True	IRT Obs
Total	Mean	42.23355	42.23174	42.31877	42.33253
	SD	18.60597	18.60003	18.70782	18.74043
	Skew	0.07469	0.07208	0.06169	0.06151
	Kurt	2.12536	2.11943	2.11758	2.11757
1 & 2	Mean	34.43899	34.43546	34.43609	34.44514
	SD	17.88142	17.87275	17.99496	18.04022
	Skew	0.44497	0.44318	0.46985	0.46703
	Kurt	2.25024	2.24348	2.29252	2.28819
3	Mean	42.93241	42.92857	43.03341	43.05253
	SD	17.21253	17.19802	17.26796	17.31328
	Skew	0.03606	0.03427	0.03909	0.04159
	Kurt	2.22310	2.21652	2.25113	2.24985
4	Mean	48.30196	48.30601	48.36802	48.37976
	SD	18.10260	18.09432	18.12650	18.16986
	Skew	-0.18644	-0.18692	-0.18429	-0.18450
	Kurt	2.22091	2.21997	2.22185	2.22256

Table 4.7 REMSD Values for Parental Education Subgroups

Exam	Parent ED Level	UEq	SEq	IRT True	IRT Obs	SDTM
Chem	1 & 2	0.03036	0.02651	0.01380	0.01353	0.01122
	3	0.02248	0.01959	0.00904	0.00847	
	4	0.02815	0.02517	0.00980	0.00960	
English	1	0.04732	0.03915	0.01988	0.01937	0.01697
	2	0.03432	0.02863	0.01704	0.01401	
	3	0.03369	0.03023	0.01000	0.00947	
	4	0.02651	0.02163	0.01060	0.00986	
French	1 & 2	0.03625	0.02898	0.01431	0.01382	0.01405
	3	0.02972	0.02315	0.00908	0.00878	
	4	0.02579	0.02209	0.00618	0.00658	
Physics	1 & 2	0.04205	0.03973	0.02505	0.02451	0.02687
	3	0.03977	0.03746	0.02809	0.02723	
	4	0.03440	0.03019	0.01578	0.01579	

Table 4.8 Old Form Cut Scores and Cumulative Proportions

		Operational		Form B	
Exam	AP Grade	Cut Score	% Below	Cut Score	% Below
Chem	2	33.5	25.72	72.5	25.60
	3	52.5	44.36	105.5	43.87
	4	76.5	67.39	133.5	67.27
	5	99.5	85.23	154.5	85.00
English	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
French	2	58.5	23.51	101.5	23.47
	3	79.5	46.17	126.5	45.87
	4	107.5	77.85	158.5	77.13
	5	126.5	92.74	178.5	92.47
Physics	2	32.5	21.77	24.5	20.87
	3	52.5	40.63	37.5	40.47
	4	81.5	67.86	52.5	67.20
	5	105.5	84.54	62.5	82.93

Table 4.9 Chemistry New Form Cut Scores

Method	Grade	Total	1 & 2	3	4
UEq	2	71.5	72.5	72.5	72.5
	3	102.5	102.5	102.5	101.5
	4	127.5	129.5	127.5	128.5
	5	148.5	146.5	148.5	148.5
IRT True	2	72.5	71.5	71.5	72.5
	3	101.5	101.5	101.5	101.5
	4	128.5	129.5	127.5	127.5
	5	149.5	150.5	148.5	148.5
IRT Observed	2	72.5	71.5	71.5	71.5
	3	101.5	102.5	101.5	101.5
	4	128.5	129.5	127.5	127.5
	5	148.5	150.5	148.5	148.5

Table 4.10 English Language New Form Cut Scores

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 4.11 French Language New Form Cut Scores

Method	Grade	Total	1 & 2	3	4
UEq	2	102.5	101.5	102.5	103.5
	3	128.5	127.5	127.5	128.5
	4	162.5	160.5	161.5	161.5
	5	182.5	183.5	181.5	182.5
IRT True	2	102.5	102.5	101.5	102.5
	3	128.5	128.5	128.5	129.5
	4	162.5	161.5	162.5	162.5
	5	181.5	180.5	181.5	182.5
IRT Observed	2	102.5	102.5	102.5	102.5
	3	128.5	128.5	128.5	129.5
	4	162.5	161.5	162.5	162.5
	5	181.5	180.5	181.5	182.5

Table 4.12 Physics B New Form Cut Scores

Method	Grade	Total	1 & 2	3	4
UEq	2	24.5	24.5	24.5	25.5
	3	37.5	37.5	37.5	37.5
	4	52.5	53.5	53.5	52.5
	5	62.5	64.5	63.5	63.5
IRT True	2	24.5	24.5	24.5	25.5
	3	37.5	37.5	37.5	37.5
	4	52.5	53.5	53.5	52.5
	5	62.5	63.5	63.5	63.5
IRT Observed	2	24.5	24.5	24.5	25.5
	3	37.5	37.5	37.5	37.5
	4	52.5	53.5	53.5	52.5
	5	62.5	63.5	63.5	63.5

Table 4.13 Classification Consistency for Parental Education Subgroups

Exam	Parent ED Level	UEq	IRT True	IRT Obs
Chem	1 & 2	96.14	98.59	98.59
	3	97.56	99.36	100
	4	98.37	98.49	98.49
English	1	99.01	100	100
	2	99.51	98.68	100
	3	98.20	100	100
	4	97.51	100	100
French	1 & 2	95.79	97.64	96.13
	3	99.38	97.70	98.48
	4	97.74	98.31	98.48
Physics	1 & 2	95.58	96.76	96.76
	3	96.76	96.76	96.76
	4	96.95	96.95	96.95

Table 4.14 Chemistry Old Form Equivalent Composite Score Moments for Unmatched Old and New Form Groups

ES	Moments	UFE	SFE	UCE	SCE	IRT True	IRT Obs
0	Mean	105.50383	105.50329	105.27664	105.29549	105.45332	105.48041
	SD	43.38026	43.35228	43.15861	43.11511	42.93200	42.89949
	Skew	-0.32199	-0.32359	-0.33049	-0.33191	-0.37405	-0.36946
	Kurt	2.08764	2.08399	2.10792	2.10253	2.13170	2.13439
0.1	Mean	104.97599	104.99037	104.63214	104.64987	104.89238	104.96096
	SD	44.00731	43.95793	43.76604	43.70530	43.20997	43.15182
	Skew	-0.28538	-0.28215	-0.28258	-0.28001	-0.33774	-0.32859
	Kurt	2.01189	2.00056	1.98723	1.97771	2.08885	2.08057
0.2	Mean	102.32235	102.34562	101.53533	101.55225	101.04161	101.08743
	SD	45.50082	45.44719	45.28460	45.23512	45.23438	45.16111
	Skew	-0.22440	-0.22235	-0.20648	-0.20339	-0.29275	-0.28845
	Kurt	1.93230	1.92442	1.92057	1.91098	1.96300	1.96304
0.3	Mean	102.26554	102.25887	101.17794	101.20719	101.13502	101.19011
	SD	44.46047	44.46546	44.45657	44.40124	44.92301	44.84776
	Skew	-0.18542	-0.18676	-0.16802	-0.16551	-0.27051	-0.26368
	Kurt	1.97219	1.97210	1.97744	1.96901	1.97693	1.97565

Table 4.15 English Language Old Form Equivalent Composite Score Moments for Unmatched Old and New Form Groups

ES	Moments	UFE	SFE	UCE	SCE	IRT True	IRT Obs
0	Mean	124.17785	124.14594	124.06333	124.02426	124.15903	124.17130
	SD	30.71220	30.79429	30.68658	30.79016	31.54858	31.44509
	Skew	-0.48235	-0.50867	-0.43668	-0.45953	-0.50521	-0.51715
	Kurt	3.16449	3.25365	3.10695	3.17909	3.26384	3.30041
0.25	Mean	124.25432	124.22423	122.48265	122.46824	122.69827	122.66270
	SD	29.16006	29.21294	29.62521	29.62220	29.01440	29.18005
	Skew	-0.43742	-0.46781	-0.50985	-0.52685	-0.38219	-0.39053
	Kurt	3.17578	3.27694	3.38598	3.42071	3.22250	3.25335
0.50	Mean	123.52249	123.59677	119.74096	119.80894	120.83874	120.76270
	SD	30.65018	30.44041	32.38934	32.18122	30.57275	30.68168
	Skew	-0.57624	-0.52445	-0.65081	-0.59452	-0.27682	-0.30159
	Kurt	3.58942	3.40712	3.55731	3.35230	3.00437	3.05604
0.75	Mean	116.30005	116.28383	110.66763	110.65455	110.83583	110.66373
	SD	32.13372	32.09100	33.03865	33.00032	30.22021	30.64619
	Skew	-0.48115	-0.49738	-0.46807	-0.47934	-0.19481	-0.20414
	Kurt	3.13155	3.14009	2.98743	2.99419	2.63883	2.66305

Table 4.16 French Language Old Form Equivalent Composite Score Moments for Unmatched Old and New Form Groups

ES	Moments	UFE	SFE	UCE	SCE	IRT True	IRT Obs
0	Mean	127.48197	127.48358	127.47013	127.47075	127.91300	127.89133
	SD	36.05744	36.02010	35.71913	35.72758	34.60584	34.65623
	Skew	-0.29830	-0.29718	-0.29445	-0.29266	-0.17755	-0.18411
	Kurt	2.66370	2.64667	2.69060	2.68966	2.51931	2.52273
0.1	Mean	126.12498	126.12165	125.74929	125.74905	125.21443	125.18951
	SD	36.10247	36.08544	35.80795	35.80431	35.60089	35.66682
	Skew	-0.23888	-0.23655	-0.21132	-0.20865	-0.12389	-0.13069
	Kurt	2.59415	2.58521	2.59346	2.59221	2.51296	2.51774
0.2	Mean	124.86688	124.86307	124.19772	124.21337	124.05913	124.01969
	SD	35.28170	35.32052	34.51344	34.52308	34.13380	34.26269
	Skew	-0.23958	-0.24010	-0.23957	-0.22967	-0.08492	-0.09416
	Kurt	2.50313	2.52215	2.54636	2.53651	2.40515	2.41137
0.3	Mean	122.70479	122.69066	121.53452	121.52046	121.10226	121.06721
	SD	35.97423	36.04046	35.75138	35.75067	35.24926	35.31038
	Skew	-0.22081	-0.22706	-0.22765	-0.23515	-0.14331	-0.15011
	Kurt	2.55116	2.58302	2.54884	2.55644	2.48362	2.49047

Table 4.17 Physics B Old Form Equivalent Composite Score Moments for Unmatched Old and New Form Groups

ES	Moments	UFE	SFE	UCE	SCE	IRT True	IRT Obs
0	Mean	43.04002	43.03840	43.08583	43.09600	43.42075	43.45012
	SD	18.69164	18.70819	18.73580	18.73211	18.75330	18.78177
	Skew	0.03948	0.03890	0.05171	0.05701	0.01042	0.01526
	Kurt	2.09610	2.10260	2.10093	2.10152	2.11504	2.11417
0.1	Mean	42.42018	42.41120	42.30173	42.29752	42.31429	42.34353
	SD	18.16642	18.17783	18.14506	18.15570	18.08577	18.14843
	Skew	0.02952	0.02685	0.05387	0.05436	0.06542	0.06768
	Kurt	2.08299	2.08901	2.09372	2.09911	2.08628	2.08037
0.2	Mean	41.51644	41.52309	41.18123	41.17440	41.16147	41.18443
	SD	18.61888	18.60977	18.66123	18.66262	18.72651	18.75065
	Skew	0.11473	0.11840	0.13517	0.13527	0.10786	0.11033
	Kurt	2.12371	2.12186	2.10123	2.10089	2.11731	2.11411
0.3	Mean	41.82788	41.83236	41.35614	41.37155	41.30934	41.33273
	SD	18.42857	18.44803	18.55026	18.54713	18.82937	18.86182
	Skew	0.08835	0.09304	0.08102	0.08971	0.07876	0.08119
	Kurt	2.12277	2.13520	2.10289	2.10326	2.00832	2.01219

Table 4.18 REMSD for Unmatched Old and New Form Groups

Exam	ES	SFE	SCE	IRT True	IRT Obs	SDTM
Chem	0.10	0.05572	0.05707	0.04608	0.04615	0.01122
	0.20	0.02839	0.02790	0.01824	0.01767	
	0.30	0.04177	0.02951	0.01803	0.01718	
English	0.25	0.12374	0.08586	0.09485	0.08733	0.01697
	0.50	0.23027	0.13212	0.15194	0.14531	
	0.75	0.33275	0.16277	0.22058	0.20359	
French	0.10	0.03262	0.03784	0.03979	0.03958	0.01405
	0.20	0.05979	0.05043	0.01885	0.01673	
	0.30	0.05053	0.04448	0.04073	0.04046	
Physics	0.10	0.04924	0.04918	0.04544	0.04329	0.02687
	0.20	0.02466	0.02090	0.02792	0.02783	
	0.30	0.06543	0.04394	0.00946	0.00861	

Table 4.19 Chemistry New Form Cut Scores for Unmatched Old and New Form Groups

Method	Grade	0	0.1	0.2	0.3
SFE	2	72.5	69.5	72.5	70.5
	3	101.5	101.5	101.5	101.5
	4	128.5	124.5	126.5	127.5
	5	148.5	147.5	147.5	146.5
SCE	2	71.5	69.5	73.5	71.5
	3	101.5	102.5	102.5	103.5
	4	128.5	124.5	127.5	127.5
	5	149.5	149.5	148.5	148.5
IRT True	2	71.5	68.5	71.5	70.5
	3	101.5	98.5	100.5	100.5
	4	128.5	126.5	128.5	127.5
	5	149.5	149.5	149.5	149.5
IRT Observed	2	70.5	68.5	71.5	70.5
	3	101.5	98.5	101.5	100.5
	4	128.5	126.5	128.5	127.5
	5	149.5	148.5	149.5	149.5

Table 4.20 English Language New Form Cut Scores for Unmatched Old and New Form Groups

Method	Grade	0	0.25	0.50	0.75
SFE	2	90.5	86.5	81.5	77.5
	3	123.5	119.5	116.5	112.5
	4	145.5	143.5	139.5	139.5
	5	163.5	162.5	158.5	159.5
SCE	2	90.5	87.5	87.5	83.5
	3	124.5	120.5	119.5	118.5
	4	145.5	146.5	141.5	144.5
	5	163.5	164.5	160.5	164.5
IRT True	2	90.5	86.5	85.5	83.5
	3	123.5	121.5	120.5	120.5
	4	144.5	144.5	141.5	145.5
	5	163.5	163.5	159.5	166.5
IRT Observed	2	90.5	86.5	85.5	83.5
	3	123.5	121.5	120.5	120.5
	4	145.5	144.5	141.5	145.5
	5	163.5	163.5	159.5	165.5

Table 4.21 French Language New Form Cut Scores for Unmatched Old and New Form Groups

Method	Grade	0	0.1	0.2	0.3
SFE	2	101.5	102.5	100.5	102.5
	3	129.5	128.5	126.5	126.5
	4	161.5	161.5	159.5	161.5
	5	180.5	182.5	182.5	178.5
SCE	2	101.5	102.5	101.5	103.5
	3	129.5	128.5	127.5	127.5
	4	160.5	162.5	160.5	162.5
	5	182.5	183.5	185.5	181.5
IRT True	2	102.5	103.5	102.5	104.5
	3	129.5	130.5	128.5	129.5
	4	162.5	163.5	161.5	161.5
	5	181.5	182.5	182.5	181.5
IRT Observed	2	102.5	103.5	102.5	104.5
	3	129.5	130.5	128.5	129.5
	4	162.5	163.5	161.5	161.5
	5	181.5	182.5	182.5	181.5

Table 4.22 Physics B New Form Cut Scores for Unmatched Old and New Form Groups

Method	Grade	0	0.1	0.2	0.3
SFE	2	24.5	23.5	24.5	23.5
	3	36.5	35.5	36.5	35.5
	4	52.5	52.5	52.5	52.5
	5	62.5	63.5	62.5	62.5
SCE	2	24.5	23.5	24.5	23.5
	3	37.5	36.5	36.5	35.5
	4	52.5	52.5	52.5	52.5
	5	62.5	63.5	62.5	62.5
IRT True	2	24.5	23.5	24.5	24.5
	3	36.5	36.5	36.5	36.5
	4	51.5	52.5	52.5	51.5
	5	62.5	63.5	63.5	62.5
IRT Observed	2	24.5	23.5	24.5	24.5
	3	36.5	36.5	36.5	36.5
	4	51.5	52.5	52.5	51.5
	5	62.5	63.5	63.5	62.5

Table 4.23 Classification Consistency with ES=0 as the Criterion

Exam	ES	SFE	SCE	IRT True	IRT Obs
Chem	0.10	97.31	96.20	96.99	96.99
	0.20	93.38	96.68	97.62	97.62
	0.30	95.09	94.26	97.61	97.61
English	0.25	89.48	91.29	95.44	94.12
	0.50	75.18	84.37	86.68	85.36
	0.75	70.11	86.97	89.66	91.59
French	0.10	93.90	94.64	94.48	94.21
	0.20	97.45	96.29	99.21	99.49
	0.30	96.22	96.80	97.79	98.30
Physics	0.10	95.39	95.32	95.09	95.09
	0.20	100	98.37	96.55	96.55
	0.30	96.98	95.35	100	100

Table 4.24 Classification Consistency with IRT True Score Equating as the Criterion

Exam	ES	SFE	SCE	IRT Obs
Chem	0	97.95	97.03	100
	0.1	95.69	96.16	100
	0.2	94.79	96.09	100
	0.3	93.67	96.36	100
English	0	98.68	97.34	98.68
	0.25	95.36	94.99	100
	0.50	89.82	97.00	100
	0.75	75.39	94.95	99.38
French	0	98.60	100	99.49
	0.1	93.71	94.61	99.22
	0.2	95.27	95.51	99.21
	0.3	96.69	96.41	100
Physics	0	98.14	96.51	100
	0.1	98.44	100	100
	0.2	98.41	98.41	100
	0.3	95.12	95.12	100

Table 4.25 Assessing the Frequency Estimation Assumption: Weighted Max Differences in Cumulative Frequency Distributions for Old and New Form Groups

Exam	ES	f_1 vs. f_2*	g_1 vs. g_2*
Chem	0	11.967	11.743
	0.1	17.888	17.186
	0.2	19.629	19.625
	0.3	23.723	23.085
English	0	16.631	15.645
	0.25	26.857	26.547
	0.50	41.870	42.247
	0.75	66.402	66.376
French	0	15.980	16.42
	0.1	16.747	16.410
	0.2	19.715	19.506
	0.3	19.559	20.148
Physics	0	13.665	15.129
	0.1	14.629	16.627
	0.2	17.596	18.528
	0.3	19.274	18.812

* $f_1(A|V)$ vs. $f_2(A|V)$ and $g_1(B|V)$ vs. $g_2(B|V)$ for all V.

Table 4.26 Assessing the Chained Equipercentile Assumption: SG REMSD Values

Exam	ES	$e_V(A)$	$e_B(V)$
Chem	0	0.03389	0.01791
	0.1	0.06897	0.05427
	0.2	0.03697	0.04459
	0.3	0.04268	0.04755
English	0	0.03871	0.04115
	0.25	0.08987	0.07768
	0.50	0.14202	0.15051
	0.75	0.15680	0.15833
French	0	0.03438	0.03441
	0.1	0.03703	0.02670
	0.2	0.04990	0.04897
	0.3	0.03956	0.04268
Physics	0	0.03961	0.03272
	0.1	0.05640	0.07060
	0.2	0.02441	0.03375
	0.3	0.06009	0.05993

Table 4.27 Reliability and Correlations of MC and FR Sections

Exam	ES	Form	Alpha MC	Alpha FR	Obs Corr	Dis Corr
Chem	0	A	0.904	0.860	0.817	0.927
		B	0.930	0.862	0.853	0.952
	0.1	A	0.906	0.864	0.827	0.935
		B	0.930	0.867	0.846	0.942
	0.2	A	0.913	0.860	0.845	0.954
		B	0.931	0.868	0.844	0.939
	0.3	A	0.910	0.869	0.848	0.953
		B	0.926	0.855	0.842	0.947
English	0	A	0.862	0.671	0.533	0.701
		B	0.867	0.679	0.546	0.711
	0.25	A	0.849	0.667	0.548	0.728
		B	0.854	0.629	0.488	0.666
	0.50	A	0.870	0.652	0.521	0.692
		B	0.849	0.619	0.438	0.604
	0.75	A	0.877	0.734	0.606	0.755
		B	0.860	0.623	0.454	0.620
French	0	A	0.908	0.855	0.748	0.849
		B	0.900	0.856	0.773	0.880
	0.1	A	0.912	0.858	0.781	0.882
		B	0.904	0.856	0.770	0.875
	0.2	A	0.905	0.853	0.782	0.890
		B	0.905	0.853	0.762	0.868
	0.3	A	0.903	0.853	0.754	0.859
		B	0.900	0.850	0.753	0.861
Physics	0	A	0.902	0.779	0.817	0.975
		B	0.910	0.746	0.793	0.962
	0.1	A	0.901	0.777	0.828	0.989
		B	0.910	0.725	0.773	0.952
	0.2	A	0.903	0.773	0.823	0.985
		B	0.907	0.731	0.790	0.971
	0.3	A	0.900	0.767	0.834	1.003
		B	0.906	0.718	0.783	0.972

Table 4.28 Chemistry Eigenvalues and Percentage of Variance Explained for the First Five Principal Components

		ES and Form							
		0		0.1		0.2		0.3	
PC		A	B	A	B	A	B	A	B
1	Eigen	17.165	21.104	17.381	21.032	18.235	21.488	17.835	20.554
	%Var	31.790	39.080	32.190	38.950	33.770	39.790	33.030	38.060
2	Eigen	1.820	1.830	1.839	1.817	1.864	1.835	1.920	1.868
	%Var	3.370	3.390	3.410	3.360	3.450	3.400	3.560	3.460
3	Eigen	1.439	1.242	1.429	1.282	1.360	1.223	1.341	1.304
	%Var	2.670	2.300	2.650	2.370	2.520	2.260	2.480	2.420
4	Eigen	1.296	1.177	1.317	1.183	1.189	1.165	1.217	1.202
	%Var	2.400	2.180	2.440	2.190	2.200	2.160	2.250	2.230
5	Eigen	1.229	1.120	1.236	1.144	1.165	1.083	1.182	1.117
	%Var	2.280	2.070	2.290	2.120	2.160	2.010	2.190	2.070

Table 4.29 English Language Eigenvalues and Percentage of Variance Explained for the First Five Principal Components

		ES and Form							
		0		0.25		0.50		0.75	
PC		A	B	A	B	A	B	A	B
1	Eigen	11.943	11.802	11.099	10.981	12.003	10.606	12.363	11.650
	%Var	27.770	27.450	25.810	25.540	27.910	24.670	28.750	27.090
2	Eigen	1.667	1.789	1.747	1.779	1.676	1.858	1.766	1.694
	%Var	3.880	4.160	4.060	4.140	3.900	4.320	4.110	3.940
3	Eigen	1.396	1.315	1.335	1.449	1.388	1.476	1.410	1.394
	%Var	3.250	3.060	3.110	3.370	3.230	3.430	3.280	3.240
4	Eigen	1.350	1.286	1.311	1.341	1.260	1.394	1.363	1.268
	%Var	3.140	2.990	3.050	3.120	2.930	3.240	3.170	2.950
5	Eigen	1.301	1.202	1.238	1.303	1.206	1.325	1.230	1.253
	%Var	3.030	2.800	2.880	3.030	2.800	3.080	2.860	2.910

Table 4.30 French Language Eigenvalues and Percentage of Variance Explained for the First Five Principal Components

		ES and Form							
		0		0.1		0.2		0.3	
PC		A	B	A	B	A	B	A	B
1	Eigen	22.810	21.685	23.468	22.409	22.311	22.347	21.840	21.736
	%Var	30.410	28.910	31.290	29.880	29.750	29.800	29.120	28.980
2	Eigen	2.769	2.955	2.733	2.896	2.864	2.705	2.730	2.950
	%Var	3.690	3.940	3.640	3.860	3.820	3.610	3.640	3.930
3	Eigen	2.397	1.988	2.251	2.070	2.276	2.143	2.261	2.178
	%Var	3.200	2.650	3.000	2.760	3.030	2.860	3.020	2.900
4	Eigen	1.737	1.633	1.625	1.556	1.641	1.607	1.550	1.595
	%Var	2.320	2.180	2.170	2.070	2.190	2.140	2.070	2.130
5	Eigen	1.520	1.475	1.419	1.444	1.415	1.418	1.448	1.454
	%Var	2.030	1.970	1.890	1.930	1.890	1.890	1.930	1.940

Table 4.31 Physics B Eigenvalues and Percentage of Variance Explained for the First Five Principal Components

		ES and Form							
		0		0.1		0.2		0.3	
PC		A	B	A	B	A	B	A	B
1	Eigen	16.236	16.764	16.080	16.599	16.288	16.381	16.045	16.363
	%Var	32.470	33.530	32.160	33.200	32.580	32.760	32.090	32.730
2	Eigen	1.932	1.848	1.943	1.908	1.896	1.939	1.851	1.910
	%Var	3.860	3.700	3.890	3.820	3.790	3.880	3.700	3.820
3	Eigen	1.447	1.421	1.416	1.343	1.400	1.315	1.425	1.496
	%Var	2.890	2.840	2.830	2.690	2.800	2.630	2.850	2.990
4	Eigen	1.315	1.219	1.219	1.251	1.349	1.225	1.342	1.217
	%Var	2.630	2.440	2.440	2.500	2.700	2.450	2.680	2.430
5	Eigen	1.193	1.149	1.189	1.173	1.145	1.179	1.182	1.142
	%Var	2.390	2.300	2.380	2.350	2.290	2.360	2.360	2.280

Table 4.32 Average and SD for Chemistry MC and FR IRT Item Parameter Estimates

Parameter	Form	ES	Ave With	Ave WO	Ave Diff	SD With	SD WO	SD Diff
MC (a)	A	0	0.88340	0.80820	0.07520	0.30782	0.28401	0.02381
		0.1	0.92720	0.84780	0.07940	0.32195	0.28807	0.03388
		0.2	0.93480	0.86420	0.07060	0.31328	0.27409	0.03919
		0.3	1.00440	0.90500	0.09940	0.35347	0.31606	0.03741
	B	0	0.95400	0.94120	0.01280	0.27659	0.27305	0.00354
		0.1	0.95600	0.93300	0.02300	0.28048	0.27895	0.00153
		0.2	0.90900	0.90260	0.00640	0.25567	0.24601	0.00966
		0.3	0.91880	0.90580	0.01300	0.30834	0.30577	0.00257
MC (b)	A	0	-0.07300	-0.17840	0.10540	0.98105	1.02798	-0.04693
		0.1	-0.00680	-0.09380	0.08700	0.88507	0.92632	-0.04125
		0.2	-0.00420	-0.05140	0.04720	0.86439	0.87231	-0.00792
		0.3	0.08300	-0.01780	0.10080	0.84402	0.91520	-0.07118
	B	0	-0.23500	-0.26020	0.02520	0.74403	0.73696	0.00707
		0.1	-0.25560	-0.30660	0.05100	0.77133	0.77088	0.00045
		0.2	-0.40200	-0.42080	0.01880	0.77006	0.75350	0.01656
		0.3	-0.47040	-0.49500	0.02460	0.80572	0.79697	0.00875
MC (c)	A	0	0.14980	0.10120	0.04860	0.10111	0.09913	0.00198
		0.1	0.16600	0.12680	0.03920	0.10602	0.09595	0.01007
		0.2	0.14340	0.11900	0.02440	0.08803	0.08825	-0.00022
		0.3	0.17200	0.12660	0.04540	0.10759	0.09593	0.01166
	B	0	0.12480	0.11360	0.01120	0.09002	0.10275	-0.01273
		0.1	0.12420	0.10640	0.01780	0.09448	0.09521	-0.00073
		0.2	0.09960	0.09320	0.00640	0.08628	0.08534	0.00094
		0.3	0.11600	0.10600	0.01000	0.10500	0.10210	0.00290
FR (a)	A	0	2.37746	2.59225	-0.21479	0.17963	0.08184	0.09779
		0.1	2.42072	2.63626	-0.21554	0.22877	0.11491	0.11386
		0.2	2.46522	2.62014	-0.15492	0.28105	0.22468	0.05637
		0.3	2.60090	2.75367	-0.15277	0.25804	0.19754	0.06050
	B	0	2.40170	2.55583	-0.15413	0.16576	0.22941	-0.06365
		0.1	2.39967	2.57556	-0.17589	0.18442	0.19749	-0.01307
		0.2	2.41016	2.59408	-0.18392	0.18328	0.27244	-0.08916
		0.3	2.28540	2.44678	-0.16137	0.15062	0.31225	-0.16163
FR (b)	A	0	0.02449	0.02959	-0.00510	1.14487	1.10986	0.03501
		0.1	0.08517	0.08845	-0.00329	1.12571	1.09743	0.02828
		0.2	0.12818	0.12970	-0.00151	1.09239	1.07199	0.02040
		0.3	0.12561	0.12631	-0.00070	1.09562	1.07644	0.01918
	B	0	-0.00718	-0.01246	0.00528	1.13869	1.12428	0.01441
		0.1	-0.05292	-0.05598	0.00306	1.11915	1.10086	0.01829
		0.2	-0.11272	-0.11294	0.00022	1.12129	1.10278	0.01851
		0.3	-0.19223	-0.18195	-0.01028	1.19171	1.17103	0.02068

Table 4.33 Average and SD for English Language MC and FR IRT Item Parameter Estimates

Parameter	Form	ES	Ave With	Ave WO	Ave Diff	SD With	SD WO	SD Diff
MC (a)	A	0	0.73050	0.72925	0.00125	0.25318	0.25607	-0.00289
		0.25	0.69875	0.69850	0.00025	0.27975	0.29908	-0.01933
		0.50	0.77225	0.76625	0.00600	0.31844	0.34332	-0.02488
		0.75	0.79075	0.78725	0.00350	0.36098	0.39384	-0.03286
	B	0	0.69900	0.69175	0.00725	0.22407	0.21984	0.00423
		0.25	0.68425	0.67875	0.00550	0.25219	0.26085	-0.00866
		0.50	0.71775	0.71275	0.00500	0.29949	0.32781	-0.02832
		0.75	0.72175	0.72175	0.00000	0.27422	0.27080	0.00342
MC (b)	A	0	-0.74050	-0.75850	0.01800	0.93469	0.92581	0.00888
		0.25	-0.75325	-0.79875	0.04550	0.97537	0.99070	-0.01533
		0.50	-0.57425	-0.59875	0.02450	0.88786	0.88808	-0.00022
		0.75	-0.20400	-0.24200	0.03800	0.79916	0.79957	-0.00041
	B	0	-0.73875	-0.76900	0.03025	0.88027	0.82715	0.05312
		0.25	-0.86200	-0.91000	0.04800	0.84141	0.86981	-0.02840
		0.50	-0.87800	-0.94375	0.06575	0.87465	0.89702	-0.02237
		0.75	-0.94675	-0.95600	0.00925	0.80135	0.80050	0.00085
MC (c)	A	0	0.12425	0.11400	0.01025	0.16049	0.16424	-0.00375
		0.25	0.11750	0.09950	0.01800	0.14397	0.15004	-0.00607
		0.50	0.11975	0.11000	0.00975	0.13916	0.13646	0.00270
		0.75	0.11800	0.10200	0.01600	0.11020	0.11882	-0.00862
	B	0	0.10750	0.09500	0.01250	0.10874	0.10751	0.00123
		0.25	0.13300	0.11075	0.02225	0.15616	0.15391	0.00225
		0.50	0.17450	0.14800	0.02650	0.18513	0.19461	-0.00948
		0.75	0.14525	0.13825	0.00700	0.16575	0.16896	-0.00321
FR (a)	A	0	1.08391	1.52747	-0.44356	0.05801	0.04568	0.01233
		0.25	1.09443	1.55025	-0.45582	0.10945	0.28901	-0.17956
		0.50	1.03493	1.50835	-0.47342	0.13861	0.35497	-0.21636
		0.75	1.27962	1.81377	-0.53415	0.05979	0.15541	-0.09562
	B	0	1.08803	1.57372	-0.48569	0.10122	0.28157	-0.18035
		0.25	0.95168	1.43211	-0.48043	0.11729	0.17720	-0.05991
		0.50	0.83830	1.37832	-0.54002	0.09938	0.10240	-0.00302
		0.75	0.90412	1.38701	-0.48289	0.11499	0.19260	-0.07761
FR (b)	A	0	-0.18811	-0.14371	-0.04440	3.21745	2.53362	0.68383
		0.25	0.00831	-0.00733	0.01564	3.29304	2.63142	0.66162
		0.50	0.10017	0.08168	0.01849	3.56218	2.79415	0.76803
		0.75	0.48017	0.37436	0.10581	2.78231	2.24823	0.53408
	B	0	-0.10110	-0.08541	-0.01569	3.17784	2.51773	0.66011
		0.25	-0.39444	-0.27656	-0.11788	3.70200	2.72328	0.97872
		0.50	-0.68976	-0.45028	-0.23948	4.13116	2.78343	1.34773
		0.75	-0.62423	-0.44101	-0.18322	3.76982	2.73449	1.03533

Table 4.34 Average and SD for French Language MC and FR IRT Item Parameter Estimates

Parameter	Form	ES	Ave With	Ave WO	Ave Diff	SD With	SD WO	SD Diff
MC (a)	A	0	0.91304	0.93679	-0.02375	0.28476	0.31082	-0.02606
		0.1	0.93464	0.91500	0.01964	0.27923	0.30424	-0.02501
		0.2	0.88643	0.91786	-0.03143	0.32107	0.33595	-0.01488
		0.3	0.92554	0.93214	-0.00661	0.36941	0.39180	-0.02239
	B	0	0.95250	0.86179	0.09071	0.37345	0.35338	0.02007
		0.1	0.93750	0.85554	0.08196	0.36126	0.34452	0.01674
		0.2	0.90786	0.84500	0.06286	0.34431	0.32075	0.02356
		0.3	0.90268	0.84607	0.05661	0.35963	0.35125	0.00838
MC (b)	A	0	-0.13214	-0.17482	0.04268	0.95665	0.97916	-0.02251
		0.1	-0.09143	-0.22661	0.13518	0.92994	1.04184	-0.11190
		0.2	-0.13161	-0.14732	0.01571	1.04495	1.05741	-0.01246
		0.3	-0.00875	-0.06357	0.05482	1.00422	1.05032	-0.04610
	B	0	-0.08196	-0.19125	0.10929	1.07036	1.14004	-0.06968
		0.1	-0.09500	-0.25661	0.16161	1.06374	1.09052	-0.02678
		0.2	-0.09161	-0.24946	0.15786	1.01440	1.09178	-0.07738
		0.3	-0.23518	-0.35929	0.12411	1.14750	1.17354	-0.02604
MC (c)	A	0	0.19411	0.17304	0.02107	0.15090	0.16053	-0.00963
		0.1	0.19018	0.13375	0.05643	0.10737	0.10812	-0.00075
		0.2	0.15786	0.14429	0.01357	0.11200	0.11989	-0.00789
		0.3	0.17732	0.14911	0.02821	0.11311	0.11865	-0.00554
	B	0	0.21179	0.15339	0.05839	0.12928	0.13409	-0.00481
		0.1	0.20643	0.14536	0.06107	0.12683	0.12661	0.00022
		0.2	0.21411	0.14893	0.06518	0.13141	0.12787	0.00354
		0.3	0.19536	0.14089	0.05446	0.13034	0.12040	0.00994
FR (a)	A	0	1.37107	1.50125	-0.13018	0.46629	0.54487	-0.07858
		0.1	1.41951	1.54130	-0.12178	0.50034	0.58866	-0.08832
		0.2	1.37105	1.48371	-0.11266	0.49437	0.57076	-0.07639
		0.3	1.36942	1.49261	-0.12319	0.49259	0.58368	-0.09109
	B	0	1.37830	1.52205	-0.14375	0.51654	0.55398	-0.03744
		0.1	1.37812	1.52862	-0.15050	0.51249	0.54628	-0.03379
		0.2	1.33930	1.48954	-0.15024	0.52991	0.59140	-0.06149
		0.3	1.31131	1.46323	-0.15192	0.49078	0.53365	-0.04287
FR (b)	A	0	-0.54525	-0.50395	-0.04129	1.79703	1.71247	0.08456
		0.1	-0.49167	-0.45436	-0.03731	1.74768	1.67141	0.07627
		0.2	-0.45087	-0.41069	-0.04018	1.86110	1.79049	0.07061
		0.3	-0.36245	-0.34333	-0.01912	1.78086	1.71421	0.06665
	B	0	-0.38684	-0.35750	-0.02933	1.58957	1.54190	0.04767
		0.1	-0.42787	-0.40734	-0.02053	1.62195	1.58656	0.03539
		0.2	-0.45999	-0.42705	-0.03294	1.66699	1.60864	0.05835
		0.3	-0.50080	-0.47691	-0.02389	1.65183	1.60663	0.04520

Table 4.35 Average and SD for Physics B MC and FR IRT Item Parameter Estimates

Parameter	Form	ES	Ave With	Ave WO	Ave Diff	SD With	SD WO	SD Diff
MC (a)	A	0	0.96717	0.94239	0.02478	0.36642	0.37076	-0.00434
		0.1	0.95087	0.94652	0.00435	0.37293	0.38025	-0.00732
		0.2	1.00370	0.97565	0.02804	0.40353	0.43031	-0.02678
		0.3	0.98043	0.95543	0.02500	0.39446	0.38075	0.01371
	B	0	0.98783	0.94109	0.04674	0.33882	0.32926	0.00956
		0.1	0.97043	0.93196	0.03848	0.32325	0.31184	0.01141
		0.2	0.93043	0.90000	0.03043	0.31237	0.30819	0.00418
		0.3	0.92065	0.89043	0.03022	0.29133	0.28108	0.01025
MC (b)	A	0	0.13500	0.09804	0.03696	0.83980	0.87071	-0.03091
		0.1	0.17913	0.16804	0.01109	0.88281	0.88268	0.00013
		0.2	0.26022	0.22000	0.04022	0.80403	0.81862	-0.01459
		0.3	0.24870	0.22783	0.02087	0.83800	0.83263	0.00537
	B	0	0.12478	0.07957	0.04522	0.77208	0.80885	-0.03677
		0.1	0.07935	0.04130	0.03804	0.80648	0.83737	-0.03089
		0.2	-0.00674	-0.04696	0.04022	0.83572	0.88433	-0.04861
		0.3	-0.03739	-0.07435	0.03696	0.82954	0.86536	-0.03582
MC (c)	A	0	0.13870	0.12174	0.01696	0.09479	0.10120	-0.00641
		0.1	0.13196	0.12565	0.00630	0.10383	0.10428	-0.00045
		0.2	0.14717	0.13065	0.01652	0.09747	0.09637	0.00110
		0.3	0.13370	0.12500	0.00870	0.09632	0.09406	0.00226
	B	0	0.13957	0.12000	0.01957	0.08132	0.08638	-0.00506
		0.1	0.13913	0.11935	0.01978	0.08993	0.09171	-0.00178
		0.2	0.13174	0.11109	0.02065	0.08567	0.08832	-0.00265
		0.3	0.14304	0.12609	0.01696	0.09917	0.09804	0.00113
FR (a)	A	0	1.93939	2.08832	-0.14893	0.31867	0.39689	-0.07822
		0.1	1.96504	2.10222	-0.13718	0.42070	0.48228	-0.06158
		0.2	1.92800	2.05551	-0.12752	0.35009	0.35671	-0.00662
		0.3	1.96509	2.03926	-0.07417	0.38736	0.35826	0.02910
	B	0	1.68685	1.84921	-0.16236	0.27846	0.28359	-0.00513
		0.1	1.55351	1.73707	-0.18356	0.25551	0.28880	-0.03329
		0.2	1.61928	1.74039	-0.12111	0.26125	0.26694	-0.00569
		0.3	1.57667	1.67325	-0.09657	0.32665	0.31125	0.01540
FR (b)	A	0	0.43152	0.40622	0.02531	1.28392	1.25268	0.03124
		0.1	0.46993	0.45988	0.01005	1.24790	1.21748	0.03042
		0.2	0.52231	0.50478	0.01753	1.28220	1.24766	0.03454
		0.3	0.53759	0.54343	-0.00584	1.25279	1.23505	0.01774
	B	0	0.56473	0.54886	0.01586	1.42809	1.35937	0.06872
		0.1	0.52831	0.50939	0.01892	1.52937	1.43686	0.09251
		0.2	0.43956	0.43067	0.00890	1.45923	1.40343	0.05580
		0.3	0.41762	0.41244	0.00518	1.52199	1.46653	0.05546

Table 4.36 Phi Coefficients for Examinee Background Variables

	Chem	English	French	Physics
N	10,941	214,049	12,188	11,186
Gender & Fee	0.04	0.02	0.01	0.05
Gender & Region	0.02	0.01	0.02	0.02
Gender & Grade	0.01	0.01	0.06	0.02
Gender & Parent ED	0.04	0.06	0.02	0.03
Gender & Ethnicity	0.07	0.05	0.03	0.08
Fee & Region	0.06	0.08	0.09	0.07
Fee & Grade	0.03	0.00	0.01	0.03
Fee & Parent Ed	0.30	0.34	0.35	0.32
Fee & Ethnicity	0.29	0.36	0.36	0.29
Region & Grade	0.17	0.20	0.20	0.11
Region & Parent Ed	0.06	0.09	0.13	0.09
Region & Ethnicity	0.27	0.29	0.32	0.30
Grade & Parent ED	0.08	0.02	0.09	0.08
Grade & Ethnicity	0.11	0.05	0.06	0.12
Parent ED & Ethnicity	0.28	0.34	0.44	0.31

Table 4.37 Statistical Significance of Background Variables in Logistic Regression Equation (Chemistry)

			Probability > Chisquare						
ES	Method	R^2	Intercept	Gender	Ethnicity	Grade	Fee	Region	Parent ED
0.1	M_2	0.0193	<0.0001	0.3427	0.1971	0.2877	0.0250	0.2906	<0.0001
	M_3	0.0007	0.3446	0.5339	0.3362	0.1346	0.1040	0.4778	--
0.2	M_2	0.0752	<0.0001	0.0519	0.1184	0.9403	0.0043	0.7295	<0.0001
	M_3	0.0028	0.2998	0.2757	0.3256	0.3538	<0.0001	0.4653	--
0.3	M_2	0.1745	<0.0001	0.1098	0.2010	0.4339	0.0462	0.6621	<0.0001
	M_3	0.0087	0.3058	0.6853	0.6707	0.4218	<0.0001	0.6736	--

Table 4.38 Statistical Significance of Background Variables in Logistic Regression Equation (English Language)

			Probability > Chisquare						
ES	Method	R^2	Intercept	Gender	Ethnicity	Grade	Fee	Region	Parent ED
0.25	M_2	0.2108	<0.0001	0.9200	0.1733	0.8362	<0.0001	0.6707	<0.0001
	M_3	0.0172	<0.0001	<0.0001	<0.0001	0.6216	<0.0001	<0.0001	--
0.50	M_2	0.5708	<0.0001	0.6773	0.1125	0.2542	0.2016	0.5257	<0.0001
	M_3	0.0603	<0.0001	<0.0001	<0.0001	0.3030	<0.0001	<0.0001	--
0.75	M_2	0.6260	<0.0001	0.0146	0.2649	0.0030	0.3084	0.3229	<0.0001
	M_3	0.1445	0.0015	<0.0001	<0.0001	0.7378	<0.0001	<0.0001	--

Table 4.39 Statistical Significance of Background Variables in Logistic Regression Equation (French Language)

			Probability > Chisquare						
ES	Method	R^2	Intercept	Gender	Ethnicity	Grade	Fee	Region	Parent ED
0.1	M_2	0.0301	<0.0001	0.3941	0.8289	0.3678	0.0041	0.9795	<0.0001
	M_3	0.0015	0.0949	0.3516	0.2297	0.0297	0.0080	0.2838	--
0.2	M_2	0.1013	<0.0001	0.4705	0.4181	0.3034	<0.0001	0.7773	<0.0001
	M_3	0.0034	0.0062	0.3513	0.3845	0.0013	0.0002	0.0746	--
0.3	M_2	0.2049	<0.0001	0.2333	0.7509	0.2480	<0.0001	0.5438	<0.0001
	M_3	0.0095	0.0013	0.1482	0.0619	<0.0001	<0.0001	0.0635	--

Table 4.40 Statistical Significance of Background Variables in Logistic Regression Equation (Physics B)

			Probability > Chisquare						
ES	Method	R^2	Intercept	Gender	Ethnicity	Grade	Fee	Region	Parent ED
0.1	M_2	0.0318	<0.0001	0.0102	0.3598	0.1134	0.0812	0.1291	<0.0001
	M_3	0.0025	0.2702	0.0161	0.7857	0.0186	0.0016	0.0581	--
0.2	M_2	0.1166	<0.0001	0.2819	0.0213	0.6396	0.0187	0.0433	<0.0001
	M_3	0.0065	0.0604	0.2572	0.9955	0.0600	<0.0001	0.0085	--
0.3	M_2	0.2678	<0.0001	0.0862	0.0375	0.4281	0.0912	0.7581	<0.0001
	M_3	0.0162	0.0879	0.1410	0.4974	0.0047	<0.0001	0.1993	--

Table 4.41 Chemistry Frequencies across Levels of Parental Education based on Matched Samples

			Parent ED			
ES	Form		1	2	3	4
0.1	A	M_0	158	279	611	452
		M_1	169	282	484	565
		M_2	161	280	500	559
		M_3	158	278	611	453
	B	M_0	136	226	388	750
		M_1	169	282	484	565
		M_2	168	280	477	575
		M_3	135	226	388	751
0.2	A	M_0	193	340	599	368
		M_1	232	232	456	580
		M_2	228	231	468	573
		M_3	183	334	608	375
	B	M_0	147	147	290	916
		M_1	232	232	456	580
		M_2	229	229	451	591
		M_3	148	149	288	915
0.3	A	M_0	211	371	597	321
		M_1	110	220	585	585
		M_2	110	220	585	585
		M_3	186	362	617	335
	B	M_0	60	121	321	998
		M_1	110	220	585	585
		M_2	110	220	585	585
		M_3	61	122	321	996

Table 4.42 English Language Frequencies across Levels of Parental Education based on Matched Samples

			Parent ED			
ES	Form		1	2	3	4
0.25	A	M_0	238	409	722	131
		M_1	89	87	1120	204
		M_2	89	87	1120	204
		M_3	202	402	757	139
	B	M_0	57	56	739	648
		M_1	89	87	1120	204
		M_2	89	87	1120	204
		M_3	59	57	739	645
0.50	A	M_0	384	659	440	17
		M_1	30	23	1394	53
		M_2	30	23	1394	53
		M_3	316	670	495	19
	B	M_0	9	7	441	1043
		M_1	30	23	1394	53
		M_2	30	23	1394	53
		M_3	10	7	448	1035
0.75	A	M_0	1270	138	46	46
		M_1	375	375	375	375
		M_2	375	375	375	375
		M_3	1200	171	64	65
	B	M_0	46	46	138	1270
		M_1	375	375	375	375
		M_2	375	375	375	375
		M_3	56	51	138	1255

Table 4.43 French Language Frequencies across Levels of Parental Education based on Matched Samples

			Parent ED			
ES	Form		1	2	3	4
0.1	A	M_0	132	203	581	584
		M_1	104	209	438	749
		M_2	104	208	439	749
		M_3	123	199	587	591
	B	M_0	81	162	342	915
		M_1	104	209	438	749
		M_2	104	208	438	750
		M_3	80	163	342	915
0.2	A	M_0	158	242	694	406
		M_1	208	208	335	749
		M_2	199	207	365	729
		M_3	146	238	703	413
	B	M_0	113	113	182	1092
		M_1	208	208	335	749
		M_2	203	203	326	768
		M_3	114	113	182	1091
0.3	A	M_0	181	278	797	244
		M_1	214	223	452	611
		M_2	213	223	454	610
		M_3	158	267	822	253
	B	M_0	85	89	181	1145
		M_1	214	223	452	611
		M_2	213	223	452	612
		M_3	89	88	180	1143

Table 4.44 Physics B Frequencies across Levels of Parental Education based on Matched Samples

			Parent ED			
ES	Form		1	2	3	4
0.1	A	M_0	166	274	612	448
		M_1	149	209	581	561
		M_2	149	209	581	561
		M_3	166	273	613	448
	B	M_0	120	167	464	749
		M_1	149	209	581	561
		M_2	149	209	581	561
		M_3	119	167	464	750
0.2	A	M_0	207	342	653	298
		M_1	194	259	528	519
		M_2	193	258	534	515
		M_3	187	338	667	308
	B	M_0	112	149	303	936
		M_1	194	259	528	519
		M_2	193	257	523	527
		M_3	112	149	305	934
0.3	A	M_0	234	387	626	253
		M_1	91	136	727	546
		M_2	91	136	727	546
		M_3	210	379	644	267
	B	M_0	42	63	337	1058
		M_1	91	136	727	546
		M_2	91	136	727	546
		M_3	42	64	338	1056

Table 4.45 Changes in ES for Matched and Unmatched Groups

Exam	Target M_0 ES	M_0 ES	M_1 ES	M_2 ES	M_3 ES
Chem	0.1	0.07507	0.01735	-0.02124	0.08591
	0.2	0.21124	0.06149	0.04432	0.21043
	0.3	0.29137	0.04107	0.02919	0.26194
English	0.25	0.23591	-0.02244	0.03984	0.16946
	0.50	0.44059	-0.03735	0.03437	0.37901
	0.75	0.72279	-0.02790	0.00258	0.54052
French	0.1	0.11761	0.05133	-0.02618	0.07821
	0.2	0.20716	0.06046	0.02541	0.18698
	0.3	0.31743	0.03196	0.01206	0.26501
Physics	0.1	0.08043	-0.00100	-0.04883	0.10622
	0.2	0.19748	0.02838	-0.01044	0.18546
	0.3	0.23373	-0.01389	0.03646	0.27105

Table 4.46 Chemistry Old Form Equivalent Composite Score Moments

		Mean		Old Form Equivalent Moments			
ES	M Method	Form B	E Method	Mean	SD	Skew	Kurtosis
0.1	M_0	108.09000	SFE	104.99037	43.95793	-0.28215	2.00056
			SCE	104.64987	43.70530	-0.28001	1.97771
			IRT True	104.89238	43.20997	-0.33774	2.08885
			IRT Obs	104.96096	43.15182	-0.32859	2.08057
	M_1	105.46333	SFE	104.80400	44.30968	-0.30813	2.01675
			SCE	104.60418	44.05745	-0.32634	2.03011
			IRT True	104.76908	43.49758	-0.35004	2.08862
			IRT Obs	104.79484	43.46116	-0.34464	2.08872
	M_2	103.78400	SFE	104.70057	44.76076	-0.29353	2.01971
			SCE	104.73701	44.71163	-0.30676	2.01943
			IRT True	105.68395	45.00945	-0.40123	2.06138
			IRT Obs	105.73707	44.92554	-0.39372	2.05709
	M_3	108.92733	SFE	105.26864	44.89668	-0.28328	1.96801
			SCE	104.96717	44.64694	-0.28590	1.95822
			IRT True	104.70553	44.19189	-0.29675	2.01322
			IRT Obs	104.74844	44.14524	-0.29117	2.01211
0.2	M_0	111.14467	SFE	102.34562	45.44719	-0.22235	1.92442
			SCE	101.55225	45.23512	-0.20339	1.91098
			IRT True	101.04161	45.23438	-0.29275	1.96300
			IRT Obs	101.08743	45.16111	-0.28845	1.96304
	M_1	104.63600	SFE	101.92613	44.76914	-0.20994	1.97643
			SCE	101.74046	44.70016	-0.19180	1.97167
			IRT True	101.91657	45.54218	-0.25115	1.94560
			IRT Obs	101.93073	45.51318	-0.24967	1.95082
	M_2	106.51400	SFE	104.45728	44.90133	-0.28755	1.96487
			SCE	104.30112	44.65807	-0.28589	1.94957
			IRT True	104.13668	44.30374	-0.36431	2.07233
			IRT Obs	104.17727	44.24448	-0.35947	2.06969
	M_3	111.14333	SFE	102.28038	45.22246	-0.21586	1.94196
			SCE	101.60994	45.09401	-0.18454	1.92798
			IRT True	100.78391	45.09960	-0.28015	1.98666
			IRT Obs	100.85111	45.01023	-0.27256	1.98271
0.3	M_0	113.90733	SFE	102.25887	44.46546	-0.18676	1.97210
			SCE	101.20719	44.40124	-0.16551	1.96901
			IRT True	101.13502	44.92301	-0.27051	1.97693
			IRT Obs	101.19011	44.84776	-0.26368	1.97565
	M_1	107.96333	SFE	106.35568	44.59831	-0.30194	2.06334
			SCE	106.14178	44.70768	-0.30382	2.04760
			IRT True	106.78333	44.94090	-0.39442	2.09345
			IRT Obs	106.82806	44.89979	-0.39055	2.09442
	M_2	108.22600	SFE	107.05309	44.40018	-0.35628	2.08144
			SCE	106.91965	44.49395	-0.35554	2.06625
			IRT True	107.16511	44.36802	-0.39536	2.12817
			IRT Obs	107.20488	44.30388	-0.39117	2.12952
	M_3	113.04800	SFE	102.34470	44.52405	-0.22494	1.96544
			SCE	101.58209	44.41845	-0.20821	1.93245
			IRT True	101.38089	44.72121	-0.26126	1.98908
			IRT Obs	101.44682	44.63688	-0.25405	1.98694

Table 4.47 English Language Old Form Equivalent Composite Score Moments

		Mean		Old Form Equivalent Moments			
ES	M Method	Form B	E Method	Mean	SD	Skew	Kurtosis
0.25	M_0	129.31467	SFE	124.22423	29.21294	-0.46781	3.27694
			SCE	122.46824	29.62220	-0.52685	3.42071
			IRT True	122.69827	29.01440	-0.38219	3.22250
			IRT Obs	122.66270	29.18005	-0.39053	3.25335
	M_1	126.73667	SFE	127.19730	28.19620	-0.45226	3.51715
			SCE	127.37705	27.23199	-0.38946	3.37558
			IRT True	127.39823	27.24820	-0.51752	3.58604
			IRT Obs	127.41651	27.30547	-0.50179	3.56181
	M_2	126.24267	SFE	125.56010	27.97380	-0.43124	3.51427
			SCE	125.42334	27.63176	-0.27092	3.26438
			IRT True	125.03771	27.08759	-0.57182	3.67748
			IRT Obs	125.02188	27.35372	-0.56015	3.66218
	M_3	129.57733	SFE	125.85973	29.54894	-0.54644	3.34772
			SCE	124.68877	28.92059	-0.52807	3.27771
			IRT True	124.70287	29.12360	-0.45361	3.30894
			IRT Obs	124.68194	29.07804	-0.45264	3.31519
0.50	M_0	133.08067	SFE	123.59677	30.44041	-0.52445	3.40712
			SCE	119.80894	32.18122	-0.59452	3.35230
			IRT True	120.83874	30.57275	-0.27682	3.00437
			IRT Obs	120.76270	30.68168	-0.30159	3.05604
	M_1	127.23667	SFE	128.03151	26.85687	-0.53323	3.52884
			SCE	128.45856	26.59672	-0.45474	3.56427
			IRT True	128.65393	26.67945	-0.44683	3.51626
			IRT Obs	128.64613	26.79584	-0.46196	3.57667
	M_2	128.04333	SFE	127.39784	27.56329	-0.45769	3.49604
			SCE	127.12966	27.14197	-0.47125	3.66852
			IRT True	127.30882	26.24386	-0.60220	3.56405
			IRT Obs	127.29668	26.57720	-0.61138	3.63950
	M_3	133.03133	SFE	125.11435	29.01908	-0.39484	3.16600
			SCE	122.32081	28.34551	-0.34323	2.96486
			IRT True	121.98216	27.58605	-0.45709	3.32986
			IRT Obs	121.90348	27.87388	-0.45905	3.34709
0.75	M_0	132.41667	SFE	116.28383	32.09100	-0.49738	3.14009
			SCE	110.65455	33.00032	-0.47934	2.99419
			IRT True	110.83583	30.22021	-0.19481	2.63883
			IRT Obs	110.66373	30.64619	-0.20414	2.66305
	M_1	121.53733	SFE	122.23119	30.58411	-0.51460	3.17385
			SCE	122.57995	30.24229	-0.45884	3.11541
			IRT True	123.04254	30.26905	-0.47547	3.20509
			IRT Obs	123.03180	30.20956	-0.47778	3.20713
	M_2	119.78733	SFE	119.87765	31.69370	-0.37947	3.09484
			SCE	119.96375	31.58707	-0.32144	3.14660
			IRT True	120.14414	31.42276	-0.40611	3.10334
			IRT Obs	120.13903	31.34240	-0.41739	3.11943
	M_3	132.46800	SFE	120.64498	30.55905	-0.42597	3.05673
			SCE	116.82947	29.71440	-0.31734	3.02424
			IRT True	115.61950	29.05728	-0.33640	3.31442
			IRT Obs	115.52572	29.24324	-0.31144	3.23425

Table 4.48 French Language Old Form Equivalent Composite Score Moments

		Mean		Old Form Equivalent Moments			
ES	M Method	Form B	E Method	Mean	SD	Skew	Kurtosis
0.1	M_0	129.89667	SFE	126.12165	36.08544	-0.23655	2.58521
			SCE	125.74905	35.80431	-0.20865	2.59221
			IRT True	125.21443	35.60089	-0.12389	2.51296
			IRT Obs	125.18951	35.66682	-0.13069	2.51774
	M_1	127.81400	SFE	126.20740	35.80455	-0.23110	2.57069
			SCE	125.92863	35.78813	-0.25565	2.59735
			IRT True	126.04668	35.01699	-0.07032	2.43776
			IRT Obs	126.00922	35.10114	-0.07818	2.44472
	M_2	127.85867	SFE	128.76859	36.43652	-0.29720	2.64532
			SCE	128.81602	36.40138	-0.27809	2.66468
			IRT True	129.33277	36.07673	-0.18082	2.52827
			IRT Obs	129.30637	36.11781	-0.18767	2.53569
	M_3	129.12667	SFE	126.69652	35.22016	-0.16589	2.61486
			SCE	126.36535	35.16396	-0.17463	2.77489
			IRT True	126.65432	34.69909	-0.13190	2.50800
			IRT Obs	126.61464	34.78125	-0.14042	2.51870
0.2	M_0	131.51200	SFE	124.86307	35.32052	-0.24010	2.52215
			SCE	124.21337	34.52308	-0.22967	2.53651
			IRT True	124.05913	34.13380	-0.08492	2.40515
			IRT Obs	124.01969	34.26269	-0.09416	2.41137
	M_1	127.49333	SFE	125.54240	37.05668	-0.24792	2.54181
			SCE	125.37433	36.88593	-0.21616	2.52062
			IRT True	125.54145	36.39839	-0.17923	2.51561
			IRT Obs	125.52307	36.44001	-0.18447	2.52236
	M_2	127.91933	SFE	127.21494	35.98360	-0.25113	2.64206
			SCE	127.22423	35.78791	-0.20019	2.68072
			IRT True	127.18180	35.95163	-0.18528	2.56517
			IRT Obs	127.18060	35.98617	-0.18857	2.56313
	M_3	131.15733	SFE	125.21937	35.91469	-0.21825	2.61908
			SCE	124.58342	35.78188	-0.20196	2.63962
			IRT True	124.27352	35.59875	-0.13811	2.52240
			IRT Obs	124.24201	35.66917	-0.14455	2.52706
0.3	M_0	132.93267	SFE	122.69066	36.04046	-0.22706	2.58302
			SCE	121.52046	35.75067	-0.23515	2.55644
			IRT True	121.10226	35.24926	-0.14331	2.48362
			IRT Obs	121.06721	35.31038	-0.15011	2.49047
	M_1	125.46800	SFE	124.33470	37.01428	-0.28300	2.61163
			SCE	124.21419	37.00975	-0.27396	2.56559
			IRT True	125.18634	36.24557	-0.08831	2.44937
			IRT Obs	125.16241	36.30382	-0.09848	2.45611
	M_2	125.73800	SFE	125.33270	36.57145	-0.28062	2.60164
			SCE	125.24723	36.36222	-0.26776	2.57538
			IRT True	125.88076	35.95338	-0.07454	2.43628
			IRT Obs	125.85753	35.99898	-0.08388	2.44011
	M_3	132.65533	SFE	124.36652	36.02955	-0.29384	2.72875
			SCE	123.35635	36.11463	-0.28736	2.82018
			IRT True	123.15939	35.08048	-0.12029	2.58821
			IRT Obs	123.14606	35.14021	-0.12346	2.59520

Table 4.49 Physics B Old Form Equivalent Composite Score Moments

		Mean		Old Form Equivalent Moments			
ES	M Method	Form B	E Method	Mean	SD	Skew	Kurtosis
0.1	M_0	43.76000	SFE	42.41120	18.17783	0.02685	2.08901
			SCE	42.29752	18.15570	0.05436	2.09911
			IRT True	42.31429	18.08577	0.06542	2.08628
			IRT Obs	42.34353	18.14843	0.06768	2.08037
	M_1	41.83667	SFE	41.90143	18.14209	-0.02574	2.09036
			SCE	41.85560	18.15809	-0.05452	2.07366
			IRT True	42.07935	18.04073	0.08077	2.08514
			IRT Obs	42.08812	18.10809	0.07810	2.08938
	M_2	41.84267	SFE	42.67502	18.59285	0.00200	2.06498
			SCE	42.75616	18.60378	0.01920	2.06722
			IRT True	42.70229	18.43910	0.06271	2.09040
			IRT Obs	42.71279	18.48383	0.06114	2.08804
	M_3	43.49400	SFE	41.82836	18.89826	0.09566	2.09196
			SCE	41.54239	19.12602	0.11200	2.10033
			IRT True	41.44611	19.16957	0.11325	2.07164
			IRT Obs	41.45683	19.19246	0.11125	2.07402
0.2	M_0	44.82867	SFE	41.52309	18.60977	0.11840	2.12186
			SCE	41.17440	18.66262	0.13527	2.10089
			IRT True	41.16147	18.72651	0.10786	2.11731
			IRT Obs	41.18443	18.75065	0.11033	2.11411
	M_1	42.03933	SFE	41.60766	18.25799	0.05863	2.10576
			SCE	41.52840	18.36966	0.07482	2.08614
			IRT True	41.52828	18.22828	0.08162	2.07496
			IRT Obs	41.54948	18.27530	0.08270	2.06742
	M_2	41.85667	SFE	42.00415	18.20239	0.05048	2.17356
			SCE	42.02754	18.11338	0.06920	2.16357
			IRT True	42.08734	18.14587	-0.00877	2.07122
			IRT Obs	42.11612	18.21044	-0.00477	2.06925
	M_3	44.13133	SFE	41.10024	18.29770	0.10039	2.14624
			SCE	40.71983	18.28890	0.11132	2.14470
			IRT True	40.73896	18.17697	0.14793	2.12275
			IRT Obs	40.75409	18.21338	0.14822	2.12283
0.3	M_0	45.63000	SFE	41.83236	18.44803	0.09304	2.13520
			SCE	41.37155	18.54713	0.08971	2.10326
			IRT True	41.30934	18.82937	0.07876	2.00832
			IRT Obs	41.33273	18.86182	0.08119	2.01219
	M_1	43.00667	SFE	43.25026	18.15557	0.02330	2.17328
			SCE	43.29236	18.16688	0.03368	2.16139
			IRT True	43.34945	18.10305	0.03081	2.11808
			IRT Obs	43.36729	18.14902	0.03121	2.11902
	M_2	43.44467	SFE	42.84832	18.25744	0.02637	2.17763
			SCE	42.79962	18.29025	0.02725	2.18980
			IRT True	43.08430	18.11117	0.04494	2.12484
			IRT Obs	43.09936	18.16182	0.04431	2.12527
	M_3	45.89333	SFE	41.37222	18.45270	0.08346	2.11024
			SCE	40.90733	18.48687	0.08181	2.10756
			IRT True	40.88370	18.26321	0.12992	2.08938
			IRT Obs	40.89357	18.28550	0.13048	2.09422

Table 4.50 Chemistry REMSD Values

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	0.05572	0.05707	0.04608	0.04615
	M_1	0.02768	0.03662	0.00946	0.00954
	M_2	0.01527	0.01386	0.02948	0.02898
	M_3	0.02769	0.02523	0.00448	0.00373
0.2	M_0	0.02839	0.02790	0.01824	0.01767
	M_1	0.03254	0.03343	0.03081	0.03082
	M_2	0.03203	0.04007	0.02519	0.02493
	M_3	0.02971	0.02089	0.02593	0.02405
0.3	M_0	0.04177	0.02951	0.01803	0.01718
	M_1	0.02629	0.02585	0.02312	0.02300
	M_2	0.02884	0.03028	0.01822	0.01753
	M_3	0.03553	0.02549	0.01829	0.01701

Note. SDTM = 0.01122.

Table 4.51 English Language REMSD Values

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.25	M_0	0.12374	0.08586	0.09485	0.08733
	M_1	0.04267	0.05274	0.04751	0.04182
	M_2	0.07553	0.09523	0.11718	0.10555
	M_3	0.10773	0.08940	0.07952	0.07800
0.50	M_0	0.23027	0.13212	0.15194	0.14531
	M_1	0.05451	0.06210	0.06199	0.05496
	M_2	0.07556	0.09388	0.10816	0.09371
	M_3	0.22392	0.15366	0.16412	0.15327
0.75	M_0	0.33275	0.16277	0.22058	0.20359
	M_1	0.04255	0.05797	0.08184	0.08001
	M_2	0.03965	0.04377	0.02952	0.02866
	M_3	0.28555	0.18082	0.17404	0.16502

Note. SDTM = 0.01697.

Table 4.52 French Language REMSD Values

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	0.03262	0.03784	0.03979	0.03958
	M_1	0.03129	0.04793	0.02039	0.02066
	M_2	0.03215	0.03558	0.01859	0.01818
	M_3	0.04463	0.04580	0.01764	0.01761
0.2	M_0	0.05979	0.05043	0.01885	0.01673
	M_1	0.04883	0.03921	0.03911	0.03820
	M_2	0.05735	0.04556	0.01951	0.01822
	M_3	0.04792	0.03902	0.02851	0.02839
0.3	M_0	0.05053	0.04448	0.04073	0.04046
	M_1	0.03161	0.05461	0.02387	0.02359
	M_2	0.03890	0.06002	0.01482	0.01438
	M_3	0.06380	0.06015	0.03394	0.03255

Note. SDTM = 0.01405.

Table 4.53 Physics B REMSD Values

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	0.04924	0.04918	0.04544	0.04329
	M_1	0.04812	0.06040	0.05651	0.05510
	M_2	0.04435	0.04735	0.06808	0.06789
	M_3	0.01876	0.02983	0.04146	0.04137
0.2	M_0	0.02466	0.02090	0.02792	0.02783
	M_1	0.06451	0.07163	0.07938	0.07868
	M_2	0.05187	0.05841	0.05553	0.05442
	M_3	0.03415	0.04937	0.04191	0.04171
0.3	M_0	0.06543	0.04394	0.00946	0.00861
	M_1	0.03895	0.03274	0.03745	0.03709
	M_2	0.03707	0.04214	0.03723	0.03650
	M_3	0.03555	0.04552	0.04221	0.04242

Note. SDTM = 0.02687.

Table 4.54 Chemistry Postsmoothed Frequency Estimation Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	72.5	101.5	128.5	148.5
0.1	M_0	69.5	101.5	124.5	147.5
	M_1	72.5	101.5	126.5	149.5
	M_2	72.5	102.5	128.5	148.5
	M_3	71.5	101.5	127.5	147.5
0.2	M_0	72.5	101.5	126.5	147.5
	M_1	70.5	103.5	129.5	148.5
	M_2	73.5	103.5	127.5	148.5
	M_3	70.5	101.5	127.5	147.5
0.3	M_0	70.5	101.5	127.5	146.5
	M_1	70.5	103.5	128.5	148.5
	M_2	70.5	102.5	128.5	148.5
	M_3	69.5	100.5	127.5	148.5

Table 4.55 Chemistry Postsmoothed Chained Equipercentile Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	71.5	101.5	128.5	149.5
0.1	M_0	69.5	102.5	124.5	149.5
	M_1	73.5	101.5	126.5	151.5
	M_2	72.5	102.5	127.5	149.5
	M_3	70.5	102.5	126.5	149.5
0.2	M_0	73.5	102.5	127.5	148.5
	M_1	70.5	104.5	129.5	148.5
	M_2	74.5	103.5	126.5	149.5
	M_3	72.5	101.5	127.5	149.5
0.3	M_0	71.5	103.5	127.5	148.5
	M_1	71.5	103.5	128.5	149.5
	M_2	71.5	102.5	127.5	148.5
	M_3	70.5	101.5	127.5	149.5

Table 4.56 Chemistry IRT True Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	71.5	101.5	128.5	149.5
0.1	M_0	68.5	98.5	126.5	149.5
	M_1	71.5	101.5	128.5	149.5
	M_2	70.5	99.5	126.5	148.5
	M_3	70.5	101.5	128.5	149.5
0.2	M_0	71.5	100.5	128.5	149.5
	M_1	72.5	101.5	128.5	149.5
	M_2	71.5	101.5	129.5	151.5
	M_3	71.5	100.5	128.5	150.5
0.3	M_0	70.5	100.5	127.5	149.5
	M_1	71.5	100.5	127.5	148.5
	M_2	71.5	101.5	127.5	149.5
	M_3	70.5	100.5	127.5	150.5

Table 4.57 Chemistry IRT Observed Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	70.5	101.5	128.5	149.5
0.1	M_0	68.5	98.5	126.5	148.5
	M_1	71.5	101.5	128.5	149.5
	M_2	70.5	99.5	126.5	148.5
	M_3	70.5	101.5	128.5	149.5
0.2	M_0	71.5	101.5	128.5	149.5
	M_1	72.5	101.5	128.5	149.5
	M_2	71.5	101.5	129.5	151.5
	M_3	71.5	101.5	128.5	150.5
0.3	M_0	70.5	100.5	127.5	149.5
	M_1	71.5	100.5	127.5	148.5
	M_2	71.5	101.5	127.5	149.5
	M_3	70.5	100.5	127.5	149.5

Table 4.58 English Language Postsmoothed Frequency Estimation Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	90.5	123.5	145.5	163.5
0.25	M_0	86.5	119.5	143.5	162.5
	M_1	88.5	124.5	145.5	162.5
	M_2	87.5	123.5	148.5	165.5
	M_3	87.5	120.5	142.5	161.5
0.50	M_0	81.5	116.5	139.5	158.5
	M_1	89.5	122.5	145.5	166.5
	M_2	84.5	124.5	145.5	163.5
	M_3	81.5	117.5	139.5	158.5
0.75	M_0	77.5	112.5	139.5	159.5
	M_1	88.5	122.5	144.5	164.5
	M_2	90.5	123.5	146.5	164.5
	M_3	81.5	113.5	139.5	158.5

Table 4.59 English Language Postsmoothed Chained Equipercentile Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	90.5	124.5	145.5	163.5
0.25	M_0	87.5	120.5	146.5	164.5
	M_1	86.5	123.5	146.5	163.5
	M_2	87.5	124.5	148.5	164.5
	M_3	88.5	121.5	143.5	164.5
0.50	M_0	87.5	119.5	141.5	160.5
	M_1	87.5	122.5	145.5	165.5
	M_2	82.5	125.5	146.5	164.5
	M_3	83.5	120.5	142.5	162.5
0.75	M_0	83.5	118.5	144.5	164.5
	M_1	87.5	122.5	144.5	163.5
	M_2	90.5	124.5	147.5	163.5
	M_3	84.5	118.5	144.5	161.5

Table 4.60 English Language IRT True Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	90.5	123.5	144.5	163.5
0.25	M_0	86.5	121.5	144.5	163.5
	M_1	88.5	123.5	146.5	165.5
	M_2	86.5	123.5	148.5	168.5
	M_3	87.5	121.5	143.5	161.5
0.50	M_0	85.5	120.5	141.5	159.5
	M_1	87.5	122.5	145.5	164.5
	M_2	85.5	122.5	146.5	167.5
	M_3	83.5	120.5	144.5	163.5
0.75	M_0	83.5	120.5	145.5	166.5
	M_1	86.5	121.5	144.5	163.5
	M_2	89.5	123.5	145.5	164.5
	M_3	83.5	121.5	145.5	164.5

Table 4.61 English Language IRT Observed Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	90.5	123.5	145.5	163.5
0.25	M_0	86.5	121.5	144.5	163.5
	M_1	88.5	123.5	146.5	164.5
	M_2	86.5	123.5	148.5	167.5
	M_3	87.5	121.5	143.5	161.5
0.50	M_0	85.5	120.5	141.5	159.5
	M_1	87.5	122.5	145.5	164.5
	M_2	86.5	122.5	146.5	166.5
	M_3	83.5	120.5	144.5	163.5
0.75	M_0	83.5	120.5	145.5	165.5
	M_1	86.5	121.5	144.5	163.5
	M_2	89.5	123.5	145.5	164.5
	M_3	83.5	121.5	145.5	164.5

Table 4.62 French Language Postsmoothed Frequency Estimation Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	101.5	129.5	161.5	180.5
0.1	M_0	102.5	128.5	161.5	182.5
	M_1	101.5	128.5	162.5	181.5
	M_2	103.5	128.5	162.5	182.5
	M_3	103.5	128.5	162.5	179.5
0.2	M_0	100.5	126.5	159.5	182.5
	M_1	104.5	128.5	160.5	182.5
	M_2	102.5	127.5	161.5	183.5
	M_3	101.5	127.5	162.5	181.5
0.3	M_0	102.5	126.5	161.5	178.5
	M_1	102.5	129.5	162.5	182.5
	M_2	101.5	128.5	162.5	183.5
	M_3	100.5	125.5	160.5	180.5

Table 4.63 French Language Postsmoothed Chained Equipercentile Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	101.5	129.5	160.5	182.5
0.1	M_0	102.5	128.5	162.5	183.5
	M_1	100.5	129.5	162.5	181.5
	M_2	102.5	129.5	162.5	182.5
	M_3	102.5	129.5	163.5	179.5
0.2	M_0	101.5	127.5	160.5	185.5
	M_1	104.5	129.5	160.5	181.5
	M_2	102.5	128.5	162.5	183.5
	M_3	101.5	129.5	162.5	182.5
0.3	M_0	103.5	127.5	162.5	181.5
	M_1	101.5	128.5	163.5	181.5
	M_2	101.5	128.5	163.5	184.5
	M_3	100.5	126.5	162.5	181.5

Table 4.64 French Language IRT True Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	102.5	129.5	162.5	181.5
0.1	M_0	103.5	130.5	163.5	182.5
	M_1	103.5	130.5	162.5	181.5
	M_2	102.5	129.5	161.5	180.5
	M_3	103.5	129.5	161.5	181.5
0.2	M_0	102.5	128.5	161.5	182.5
	M_1	103.5	129.5	161.5	181.5
	M_2	101.5	128.5	162.5	182.5
	M_3	103.5	130.5	163.5	182.5
0.3	M_0	104.5	129.5	161.5	181.5
	M_1	103.5	129.5	161.5	180.5
	M_2	102.5	129.5	162.5	181.5
	M_3	102.5	128.5	161.5	181.5

Table 4.65 French Language IRT Observed Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	102.5	129.5	162.5	181.5
0.1	M_0	103.5	130.5	163.5	182.5
	M_1	103.5	130.5	162.5	181.5
	M_2	102.5	129.5	161.5	181.5
	M_3	103.5	129.5	161.5	181.5
0.2	M_0	102.5	128.5	161.5	182.5
	M_1	103.5	129.5	161.5	181.5
	M_2	101.5	128.5	162.5	182.5
	M_3	103.5	130.5	163.5	182.5
0.3	M_0	104.5	129.5	161.5	181.5
	M_1	103.5	129.5	161.5	180.5
	M_2	102.5	129.5	162.5	181.5
	M_3	102.5	128.5	161.5	181.5

Table 4.66 Physics B Postsmoothed Frequency Estimation Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	24.5	36.5	52.5	62.5
0.1	M_0	23.5	35.5	52.5	63.5
	M_1	24.5	36.5	53.5	63.5
	M_2	25.5	37.5	52.5	63.5
	M_3	24.5	36.5	52.5	62.5
0.2	M_0	24.5	36.5	52.5	62.5
	M_1	24.5	36.5	53.5	65.5
	M_2	24.5	36.5	53.5	64.5
	M_3	24.5	36.5	52.5	63.5
0.3	M_0	23.5	35.5	52.5	62.5
	M_1	25.5	36.5	52.5	63.5
	M_2	25.5	36.5	53.5	63.5
	M_3	24.5	36.5	52.5	62.5

Table 4.67 Physics B Postsmoothed Chained Equipercentile Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	24.5	37.5	52.5	62.5
0.1	M_0	23.5	36.5	52.5	63.5
	M_1	24.5	36.5	52.5	64.5
	M_2	25.5	37.5	52.5	63.5
	M_3	24.5	37.5	52.5	62.5
0.2	M_0	24.5	36.5	52.5	62.5
	M_1	25.5	36.5	53.5	65.5
	M_2	24.5	36.5	53.5	64.5
	M_3	25.5	36.5	53.5	64.5
0.3	M_0	23.5	35.5	52.5	62.5
	M_1	24.5	36.5	53.5	63.5
	M_2	24.5	36.5	53.5	63.5
	M_3	24.5	36.5	52.5	63.5

Table 4.68 Physics B IRT True Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	24.5	36.5	51.5	62.5
0.1	M_0	23.5	36.5	52.5	63.5
	M_1	24.5	37.5	53.5	63.5
	M_2	24.5	37.5	53.5	63.5
	M_3	24.5	37.5	52.5	62.5
0.2	M_0	24.5	36.5	52.5	63.5
	M_1	24.5	37.5	53.5	64.5
	M_2	24.5	37.5	53.5	64.5
	M_3	24.5	37.5	52.5	63.5
0.3	M_0	24.5	36.5	51.5	62.5
	M_1	24.5	37.5	52.5	63.5
	M_2	24.5	37.5	52.5	63.5
	M_3	24.5	37.5	52.5	63.5

Table 4.69 Physics B IRT Observed Score Cut Scores

		Grade			
ES	Match Method	2	3	4	5
0	M_0	24.5	36.5	51.5	62.5
0.1	M_0	23.5	36.5	52.5	63.5
	M_1	24.5	37.5	53.5	63.5
	M_2	24.5	37.5	53.5	63.5
	M_3	24.5	37.5	52.5	62.5
0.2	M_0	24.5	36.5	52.5	63.5
	M_1	24.5	37.5	53.5	64.5
	M_2	24.5	37.5	53.5	64.5
	M_3	24.5	37.5	52.5	63.5
0.3	M_0	24.5	36.5	51.5	62.5
	M_1	24.5	37.5	52.5	63.5
	M_2	24.5	37.5	52.5	63.5
	M_3	24.5	37.5	52.5	62.5

Table 4.70 Chemistry Classification Consistency with ES=0 as the Criterion

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	93.90	94.64	94.48	94.21
	M_1	97.53	95.43	100	99.49
	M_2	99.27	97.74	95.51	96.03
	M_3	97.62	97.06	99.49	100
0.2	M_0	97.45	96.29	99.21	99.49
	M_1	96.41	95.46	99.38	98.87
	M_2	96.90	95.07	97.48	96.97
	M_3	97.11	98.48	98.41	98.69
0.3	M_0	96.22	96.80	97.79	98.30
	M_1	97.36	98.48	97.53	97.01
	M_2	98.14	97.58	99.09	98.58
	M_3	96.63	98.58	96.99	98.30

Table 4.71 English Language Classification Consistency with ES=0 as the Criterion

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

Table 4.72 French Language Classification Consistency with ES=0 as the Criterion

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	97.31	96.20	96.99	96.99
	M_1	97.62	97.04	98.35	98.35
	M_2	95.68	97.57	98.59	99.13
	M_3	96.20	95.11	98.37	98.37
0.2	M_0	93.38	96.68	97.62	97.62
	M_1	94.87	97.30	98.37	98.37
	M_2	95.88	96.20	97.85	97.85
	M_3	96.58	98.21	96.99	96.99
0.3	M_0	95.09	94.26	97.61	97.61
	M_1	97.41	95.87	97.83	97.83
	M_2	96.69	95.51	100	100
	M_3	94.36	93.97	98.16	98.16

Table 4.73 Physics B Classification Consistency with ES=0 as the Criterion

		Equating Method			
ES	Method	SFE	SCE	IRT True	IRT Obs
0.1	M_0	95.39	95.32	95.09	95.09
	M_1	96.76	95.60	93.27	93.27
	M_2	95.32	96.95	93.27	93.27
	M_3	100	100	96.51	96.51
0.2	M_0	100	98.37	96.55	96.55
	M_1	94.38	91.29	92.09	92.09
	M_2	95.58	93.95	92.09	92.09
	M_3	98.41	92.49	94.92	94.92
0.3	M_0	96.98	95.35	100	100
	M_1	96.95	95.13	94.92	94.92
	M_2	95.30	95.13	94.92	94.92
	M_3	100	96.78	94.92	96.51

Table 4.74 Chemistry Classification Consistency with IRT True Score Equating Relationship as the Criterion

		Equating Method		
ES	Method	SFE	SCE	IRT Obs
0	M_0	98.60	100	99.49
0.1	M_0	93.71	94.61	99.22
	M_1	97.69	95.43	100
	M_2	94.95	95.08	100
	M_3	96.95	97.57	100
0.2	M_0	95.27	95.51	99.21
	M_1	95.64	94.84	100
	M_2	92.98	92.55	100
	M_3	95.36	96.89	99.21
0.3	M_0	96.69	96.41	100
	M_1	96.28	96.01	100
	M_2	97.07	98.49	100
	M_3	97.87	98.41	99.20

Table 4.75 English Language Classification Consistency with IRT True Score Equating Relationship as the Criterion

		Equating Method		
ES	Method	SFE	SCE	IRT Obs
0	M_0	98.68	97.34	98.68
0.25	M_0	95.36	94.99	100
	M_1	95.45	97.83	99.43
	M_2	97.88	95.97	99.49
	M_3	97.39	97.41	100
0.50	M_0	89.82	97.00	100
	M_1	97.85	99.43	100
	M_2	93.24	93.01	99.01
	M_3	84.87	96.69	100
0.75	M_0	75.39	94.95	99.38
	M_1	97.07	98.23	100
	M_2	98.23	94.96	100
	M_3	76.81	92.32	100

Table 4.76 French Language Classification Consistency with IRT True Score Equating Relationship as the Criterion

		Equating Method		
ES	Method	SFE	SCE	IRT Obs
0	M_0	97.95	97.03	100
0.1	M_0	95.69	96.16	100
	M_1	96.74	97.08	100
	M_2	96.32	98.05	99.46
	M_3	97.07	96.45	100
0.2	M_0	94.79	96.09	100
	M_1	96.81	98.32	100
	M_2	97.06	98.96	100
	M_3	94.34	96.89	100
0.3	M_0	93.67	96.36	100
	M_1	97.29	95.39	100
	M_2	97.45	96.12	100
	M_3	94.15	95.76	100

Table 4.77 Physics B Classification Consistency with IRT True Score Equating Relationship as the Criterion

		Equating Method		
ES	Method	SFE	SCE	IRT Obs
0	M_0	98.14	96.51	100
0.1	M_0	98.44	100	100
	M_1	98.37	95.54	100
	M_2	96.89	96.89	100
	M_3	98.37	100	100
0.2	M_0	98.41	98.41	100
	M_1	97.17	95.71	100
	M_2	98.37	98.37	100
	M_3	98.37	94.08	100
0.3	M_0	95.12	95.12	100
	M_1	96.92	96.72	100
	M_2	95.26	96.72	100
	M_3	96.78	98.37	98.41

Table 4.78 Chemistry Weighted Max Difference in Cumulative Frequency Distributions

ES	Method	f_1 vs. f_2*	g_1 vs. g_2*
0	M_0	11.967	11.743
0.1	M_0	17.888	17.186
	M_1	16.821	17.008
	M_2	16.213	16.830
	M_3	14.696	15.408
0.2	M_0	19.629	19.625
	M_1	16.487	15.567
	M_2	17.437	16.535
	M_3	22.576	21.818
0.3	M_0	23.723	23.085
	M_1	15.748	16.008
	M_2	15.671	17.005
	M_3	24.144	22.608

* $f_1(A|V)$ vs. $f_2(A|V)$ and $g_1(B|V)$ vs. $g_2(B|V)$ for all V.

Table 4.79 English Language Weighted Max Difference in Cumulative Frequency Distributions

ES	Method	f_1 vs. f_2*	g_1 vs. g_2*
0	M_0	16.631	15.645
0.25	M_0	26.857	26.547
	M_1	17.323	18.135
	M_2	20.764	20.889
	M_3	25.667	24.904
0.50	M_0	41.870	42.247
	M_1	19.294	19.693
	M_2	14.540	15.549
	M_3	38.541	39.311
0.75	M_0	66.402	66.376
	M_1	21.506	19.650
	M_2	19.037	19.207
	M_3	56.566	58.507

* $f_1(A|V)$ vs. $f_2(A|V)$ and $g_1(B|V)$ vs. $g_2(B|V)$ for all V.

Table 4.80 French Language Weighted Max Difference in Cumulative Frequency Distributions

ES	Method	f_1 vs. f_2*	g_1 vs. g_2*
0	M_0	15.980	16.420
0.1	M_0	16.747	16.410
	M_1	13.259	12.663
	M_2	14.692	15.274
	M_3	16.476	15.269
0.2	M_0	19.715	19.506
	M_1	17.638	17.456
	M_2	14.712	15.569
	M_3	20.107	20.555
0.3	M_0	19.559	20.148
	M_1	15.771	15.699
	M_2	16.160	16.532
	M_3	22.573	21.834

* $f_1(A|V)$ vs. $f_2(A|V)$ and $g_1(B|V)$ vs. $g_2(B|V)$ for all V.

Table 4.81 Physics B Weighted Max Difference in Cumulative Frequency Distributions

ES	Method	f_1 vs. f_2*	g_1 vs. g_2*
0	M_0	13.665	15.129
0.1	M_0	14.629	16.627
	M_1	17.192	17.458
	M_2	13.961	13.467
	M_3	17.002	15.815
0.2	M_0	17.596	18.528
	M_1	14.859	14.198
	M_2	12.941	13.366
	M_3	18.774	19.932
0.3	M_0	19.274	18.812
	M_1	14.140	14.834
	M_2	14.951	15.627
	M_3	19.751	21.735

* $f_1(A|V)$ vs. $f_2(A|V)$ and $g_1(B|V)$ vs. $g_2(B|V)$ for all V.

Table 4.82 Assessing the Chained Equipercentile Assumption: Chemistry SG REMSD Values

ES	Method	$e_V(A)$	$e_B(V)$
0	M_0	0.03389	0.01791
0.1	M_0	0.06897	0.05427
	M_1	0.04220	0.04187
	M_2	0.03575	0.02169
	M_3	0.03231	0.02809
0.2	M_0	0.03697	0.04459
	M_1	0.03508	0.03715
	M_2	0.03928	0.03157
	M_3	0.03877	0.03106
0.3	M_0	0.04268	0.04755
	M_1	0.03050	0.02509
	M_2	0.03189	0.03131
	M_3	0.04926	0.02937

Table 4.83 Assessing the Chained Equipercentile Assumption: English Language SG REMSD Values

ES	Method	$e_V(A)$	$e_B(V)$
0	M_0	0.03871	0.04115
0.25	M_0	0.08987	0.07768
	M_1	0.03715	0.04400
	M_2	0.09457	0.09854
	M_3	0.08097	0.07766
0.50	M_0	0.14202	0.15051
	M_1	0.04204	0.05142
	M_2	0.08565	0.09613
	M_3	0.13728	0.14658
0.75	M_0	0.15680	0.15833
	M_1	0.04415	0.05252
	M_2	0.05081	0.05571
	M_3	0.15762	0.17087

Table 4.84 Assessing the Chained Equipercentile Assumption: French Language SG REMSD Values

ES	Method	$e_V(A)$	$e_B(V)$
0	M_0	0.03438	0.03441
0.1	M_0	0.03703	0.02670
	M_1	0.04107	0.04535
	M_2	0.03732	0.03057
	M_3	0.04518	0.04808
0.2	M_0	0.04990	0.04897
	M_1	0.03915	0.02542
	M_2	0.03731	0.03603
	M_3	0.03992	0.03815
0.3	M_0	0.03956	0.04268
	M_1	0.04930	0.05132
	M_2	0.05377	0.06745
	M_3	0.06275	0.05928

Table 4.85 Assessing the Chained Equipercentile Assumption: Physics B SG REMSD Values

ES	Method	$e_V(A)$	$e_B(V)$
0	M_0	0.03961	0.03272
0.1	M_0	0.05640	0.07060
	M_1	0.04657	0.05335
	M_2	0.02187	0.03942
	M_3	0.03885	0.04403
0.2	M_0	0.02441	0.03375
	M_1	0.04760	0.03391
	M_2	0.04253	0.03401
	M_3	0.03937	0.04326
0.3	M_0	0.06009	0.05993
	M_1	0.03335	0.03165
	M_2	0.04111	0.03459
	M_3	0.05111	0.05196

Table 4.86 Minimum and Maximum Observed and Disattenuated Correlations

	Observed Correlation		Disattenuated Correlation	
Exam	Min	Max	Min	Max
Chemistry	0.817	0.866	0.926	0.961
English	0.438	0.606	0.604	0.755
French	0.739	0.786	0.848	0.891
Physics	0.768	0.834	0.946	1.003

Table 4.87 Chemistry Average MC Item Parameter Estimates

		Item Parameter Estimates								
		a			b			c		
Form	Match	MC+FR	MC	Diff	MC+FR	MC	Diff	MC+FR	MC	Diff
A	M_0	0.92720	0.84780	0.07940	-0.00680	-0.09380	0.08700	0.16600	0.12680	0.03920
		0.93480	0.86420	0.07060	-0.00420	-0.05140	0.04720	0.14340	0.11900	0.02440
		1.00440	0.90500	0.09940	0.08300	-0.01780	0.10080	0.17200	0.12660	0.04540
	M_1	0.93500	0.84900	0.08600	-0.05420	-0.16340	0.10920	0.15840	0.11280	0.04560
		0.96560	0.90020	0.06540	0.00560	-0.03700	0.04260	0.15740	0.13600	0.02140
		0.94980	0.88060	0.06920	-0.08860	-0.15500	0.06640	0.15240	0.12200	0.03040
	M_2	0.99760	0.89960	0.09800	0.00280	-0.11980	0.12260	0.17400	0.11840	0.05560
		0.95260	0.86860	0.08400	-0.03000	-0.10860	0.07860	0.16680	0.13500	0.03180
		0.90840	0.85680	0.05160	-0.12240	-0.18420	0.06180	0.14740	0.11980	0.02760
	M_3	0.93220	0.86660	0.06560	-0.03300	-0.12440	0.09140	0.16720	0.12940	0.03780
		0.97960	0.89140	0.08820	0.07000	-0.01520	0.08520	0.17460	0.13940	0.03520
		0.96620	0.89360	0.07260	0.05280	-0.02180	0.07460	0.16260	0.13120	0.03140
B	M_0	0.95600	0.93300	0.02300	-0.25560	-0.30660	0.05100	0.12420	0.10640	0.01780
		0.90900	0.90260	0.00640	-0.40200	-0.42080	0.01880	0.09960	0.09320	0.00640
		0.91880	0.90580	0.01300	-0.47040	-0.49500	0.02460	0.11600	0.10600	0.01000
	M_1	0.98000	0.96380	0.01620	-0.22620	-0.26740	0.04120	0.12800	0.11220	0.01580
		0.97980	0.94420	0.03560	-0.20940	-0.26920	0.05980	0.12200	0.09580	0.02620
		0.93480	0.93140	0.00340	-0.31420	-0.33540	0.02120	0.11340	0.10800	0.00540
	M_2	0.97140	0.96820	0.00320	-0.18340	-0.20960	0.02620	0.12140	0.11060	0.01080
		0.99140	0.96900	0.02240	-0.25100	-0.30000	0.04900	0.11620	0.09880	0.01740
		0.92540	0.91200	0.01340	-0.31720	-0.35660	0.03940	0.10940	0.09440	0.01500
	M_3	0.95040	0.93560	0.01480	-0.29700	-0.34400	0.04700	0.12660	0.10880	0.01780
		0.91380	0.90260	0.01120	-0.40060	-0.43880	0.03820	0.10540	0.08960	0.01580
		0.87960	0.86420	0.01540	-0.46220	-0.50040	0.03820	0.10640	0.08980	0.01660

Table 4.88 Chemistry Average FR Item Parameter Estimates

		Item Parameter Estimates					
		a			b		
Form	Match	MC+FR	FR	Diff	MC+FR	FR	Diff
A	M_0	2.42072	2.63626	-0.21554	0.08517	0.08845	-0.00329
		2.46522	2.62014	-0.15492	0.12818	0.12970	-0.00151
		2.60090	2.75367	-0.15277	0.12561	0.12631	-0.00070
	M_1	2.40383	2.61995	-0.21612	0.04976	0.05410	-0.00434
		2.48515	2.64801	-0.16286	0.07479	0.07726	-0.00246
		2.46413	2.62308	-0.15895	-0.00710	-0.00166	-0.00544
	M_2	2.44337	2.63710	-0.19374	0.04392	0.04892	-0.00500
		2.45454	2.59110	-0.13656	0.02916	0.03181	-0.00266
		2.37760	2.51503	-0.13743	-0.02246	-0.01937	-0.00309
	M_3	2.45584	2.63682	-0.18098	0.05415	0.05698	-0.00283
		2.54730	2.69939	-0.15209	0.12377	0.12559	-0.00182
		2.47825	2.65760	-0.17935	0.11304	0.11210	0.00094
B	M_0	2.39967	2.57556	-0.17589	-0.05292	-0.05598	0.00306
		2.41016	2.59408	-0.18392	-0.11272	-0.11294	0.00022
		2.28540	2.44678	-0.16137	-0.19223	-0.18195	-0.01028
	M_1	2.46070	2.64448	-0.18379	0.04738	0.04103	0.00635
		2.52384	2.68680	-0.16296	0.05659	0.05474	0.00185
		2.41107	2.60314	-0.19207	-0.02735	-0.02762	0.00026
	M_2	2.48648	2.65561	-0.16914	0.05701	0.05072	0.00629
		2.49944	2.65463	-0.15519	0.01263	0.00717	0.00547
		2.41954	2.58191	-0.16237	-0.03715	-0.03701	-0.00014
	M_3	2.45528	2.62465	-0.16937	-0.03723	-0.04088	0.00365
		2.39347	2.58676	-0.19329	-0.10354	-0.09773	-0.00581
		2.32791	2.44529	-0.11737	-0.16234	-0.15339	-0.00895

Table 4.89 English Language Average MC Item Parameter Estimates

		Item Parameter Estimates								
		a			b			c		
Form	Match	MC+FR	MC	Diff	MC+FR	MC	Diff	MC+FR	MC	Diff
A	M_0	0.69875	0.69850	0.00025	-0.75325	-0.79875	0.04550	0.11750	0.09950	0.01800
		0.77225	0.76625	0.00600	-0.57425	-0.59875	0.02450	0.11975	0.11000	0.00975
		0.79075	0.78725	0.00350	-0.20400	-0.24200	0.03800	0.11800	0.10200	0.01600
	M_1	0.70500	0.69150	0.01350	-0.80125	-0.87000	0.06875	0.17950	0.16050	0.01900
		0.65500	0.64875	0.00625	-1.03750	-1.11300	0.07550	0.12800	0.10750	0.02050
		0.71450	0.72775	-0.01325	-0.73350	-0.68100	-0.05250	0.11600	0.12625	-0.01025
	M_2	0.74975	0.74225	0.00750	-0.83750	-0.87875	0.04125	0.13100	0.11325	0.01775
		0.61775	0.60700	0.01075	-1.01225	-1.05950	0.04725	0.09800	0.08225	0.01575
		0.73375	0.73975	-0.00600	-0.64675	-0.61125	-0.03550	0.12475	0.12575	-0.00100
	M_3	0.66000	0.65975	0.00025	-0.86325	-0.89750	0.03425	0.10325	0.08800	0.01525
		0.73250	0.70500	0.02750	-0.56975	-0.63675	0.06700	0.13650	0.11400	0.02250
		0.73950	0.74725	-0.00775	-0.47700	-0.44100	-0.03600	0.09950	0.10650	-0.00700
B	M_0	0.68425	0.67875	0.00550	-0.86200	-0.91000	0.04800	0.13300	0.11075	0.02225
		0.71775	0.71275	0.00500	-0.87800	-0.94375	0.06575	0.17450	0.14800	0.02650
		0.72175	0.72175	0.00000	-0.94675	-0.95600	0.00925	0.14525	0.13825	0.00700
	M_1	0.67425	0.67650	-0.00225	-0.78400	-0.80425	0.02025	0.11925	0.11300	0.00625
		0.66275	0.66425	-0.00150	-0.76500	-0.80100	0.03600	0.13900	0.12850	0.01050
		0.73650	0.73125	0.00525	-0.62025	-0.65425	0.03400	0.10400	0.08825	0.01575
	M_2	0.71475	0.71700	-0.00225	-0.83000	-0.85000	0.02000	0.09425	0.08575	0.00850
		0.65225	0.65550	-0.00325	-0.89200	-0.88250	-0.00950	0.09825	0.09150	0.00675
		0.76350	0.75775	0.00575	-0.50850	-0.56625	0.05775	0.11925	0.09650	0.02275
	M_3	0.67575	0.68250	-0.00675	-0.91050	-0.90825	-0.00225	0.09350	0.08850	0.00500
		0.70650	0.70525	0.00125	-0.98250	-0.96625	-0.01625	0.11500	0.11650	-0.00150
		0.69675	0.70100	-0.00425	-1.08075	-1.09300	0.01225	0.07325	0.06725	0.00600

Table 4.90 English Language Average FR Item Parameter Estimates

		Item Parameter Estimates					
		a			b		
Form	Match	MC+FR	FR	Diff	MC+FR	FR	Diff
A	M_0	1.09443	1.55025	-0.45581	0.00831	-0.00733	0.01564
		1.03493	1.50835	-0.47342	0.10017	0.08168	0.01849
		1.27962	1.81377	-0.53415	0.48017	0.37436	0.10581
	M_1	0.97371	1.40929	-0.43558	-0.37337	-0.26158	-0.11179
		0.91274	1.45564	-0.54290	-0.42476	-0.29198	-0.13278
		1.03050	1.60927	-0.57877	0.05247	0.02251	0.02996
	M_2	1.03816	1.56617	-0.52801	-0.22324	-0.16729	-0.05596
		1.02972	1.51061	-0.48089	-0.30275	-0.23227	-0.07049
		1.08025	1.62944	-0.54919	0.00693	0.00098	0.00595
	M_3	1.01823	1.46080	-0.44257	-0.07934	-0.06122	-0.01813
		1.06446	1.50307	-0.43861	0.09684	0.06162	0.03522
		1.11155	1.70549	-0.59393	0.38328	0.27759	0.10569
B	M_0	0.95168	1.43211	-0.48043	-0.39444	-0.27656	-0.11788
		0.83830	1.37832	-0.54002	-0.68976	-0.45028	-0.23948
		0.90412	1.38701	-0.48289	-0.62423	-0.44101	-0.18323
	M_1	0.95211	1.49982	-0.54772	-0.26848	-0.20394	-0.06454
		0.86907	1.38050	-0.51143	-0.35068	-0.24719	-0.10349
		1.08798	1.57191	-0.48393	0.05067	0.03458	0.01609
	M_2	0.88900	1.40920	-0.52020	-0.33052	-0.24043	-0.09009
		0.89637	1.37936	-0.48299	-0.37061	-0.25654	-0.11407
		1.15582	1.65379	-0.49797	0.09702	0.07841	0.01861
	M_3	1.02980	1.45192	-0.42212	-0.49170	-0.38529	-0.10641
		0.94715	1.46395	-0.51680	-0.66307	-0.47185	-0.19122
		1.00407	1.49829	-0.49422	-0.54165	-0.40753	-0.13412

Table 4.91 French Language Average MC Item Parameter Estimates

		Item Parameter Estimates								
		a			b			c		
Form	Match	MC+FR	MC	Diff	MC+FR	MC	Diff	MC+FR	MC	Diff
A	M_0	0.93464	0.91500	0.01964	-0.09143	-0.22661	0.13518	0.19018	0.13375	0.05643
		0.88643	0.91786	-0.03143	-0.13161	-0.14732	0.01571	0.15786	0.14429	0.01357
		0.92554	0.93214	-0.00661	-0.00875	-0.06357	0.05482	0.17732	0.14911	0.02821
	M_1	0.88893	0.89268	-0.00375	-0.15214	-0.20446	0.05232	0.18179	0.15625	0.02554
		0.91946	0.93393	-0.01446	-0.16804	-0.20804	0.04000	0.16161	0.14179	0.01982
		0.89054	0.92536	-0.03482	-0.09643	-0.13161	0.03518	0.17196	0.15732	0.01464
	M_2	0.90804	0.91964	-0.01161	-0.18625	-0.26446	0.07821	0.18679	0.15893	0.02786
		0.96768	0.96661	0.00107	-0.14143	-0.22268	0.08125	0.18857	0.15357	0.03500
		0.87964	0.90554	-0.02589	-0.15821	-0.18679	0.02857	0.16500	0.15393	0.01107
	M_3	0.88179	0.90946	-0.02768	-0.19000	-0.23786	0.04786	0.16964	0.14589	0.02375
		0.94161	0.95000	-0.00839	-0.08321	-0.13018	0.04696	0.17982	0.15607	0.02375
		0.89482	0.92982	-0.03500	-0.07304	-0.11161	0.03857	0.16911	0.15107	0.01804
B	M_0	0.93750	0.85554	0.08196	-0.09500	-0.25661	0.16161	0.20643	0.14536	0.06107
		0.90786	0.84500	0.06286	-0.09161	-0.24946	0.15786	0.21411	0.14893	0.06518
		0.90268	0.84607	0.05661	-0.23518	-0.35929	0.12411	0.19536	0.14089	0.05446
	M_1	0.93786	0.85857	0.07929	-0.04268	-0.16607	0.12339	0.21054	0.15893	0.05161
		0.94054	0.87518	0.06536	-0.05571	-0.15607	0.10036	0.19304	0.14732	0.04571
		0.96036	0.87161	0.08875	0.01875	-0.13464	0.15339	0.21179	0.14393	0.06786
	M_2	0.93089	0.84643	0.08446	0.02393	-0.15321	0.17714	0.22857	0.15929	0.06929
		0.95554	0.88375	0.07179	-0.02089	-0.12196	0.10107	0.20875	0.16696	0.04179
		0.95214	0.86393	0.08821	0.00268	-0.13304	0.13571	0.20893	0.14446	0.06446
	M_3	0.91125	0.86375	0.04750	-0.03714	-0.14625	0.10911	0.22661	0.17643	0.05018
		0.89625	0.84036	0.05589	-0.17732	-0.29714	0.11982	0.19679	0.14554	0.05125
		0.83250	0.78786	0.04464	-0.24589	-0.38518	0.13929	0.18821	0.12786	0.06036

Table 4.92 French Language Average FR Item Parameter Estimates

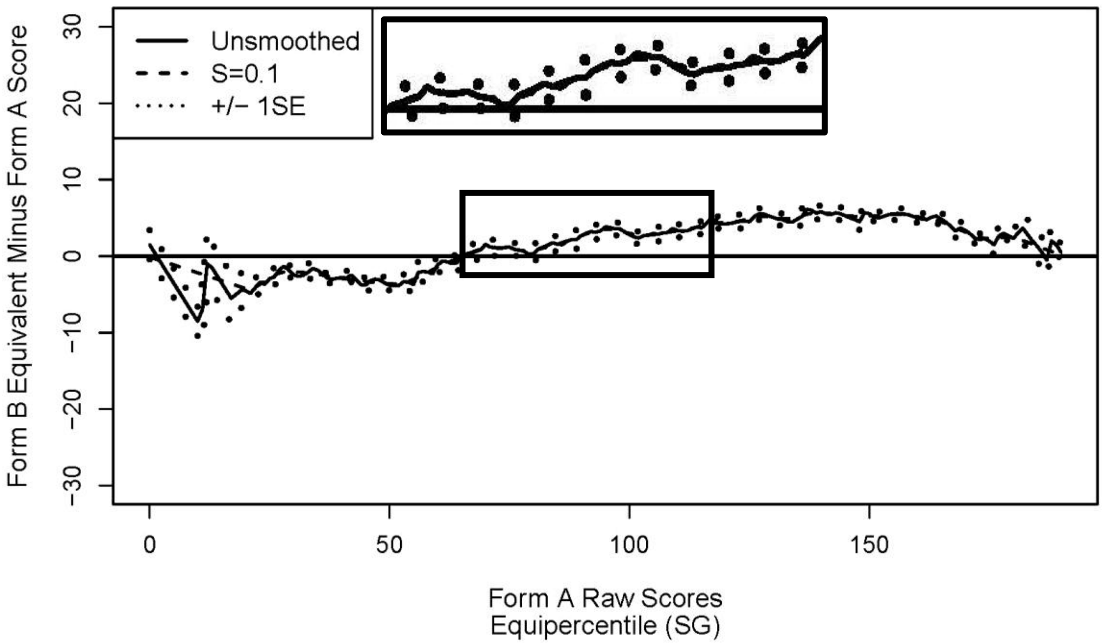
		Item Parameter Estimates					
		a			b		
Form	Match	MC+FR	FR	Diff	MC+FR	FR	Diff
A	M_0	1.41951	1.54130	-0.12178	-0.49167	-0.45436	-0.03731
		1.37105	1.48371	-0.11266	-0.45087	-0.41069	-0.04018
		1.36942	1.49261	-0.12319	-0.36245	-0.34333	-0.01912
	M_1	1.39051	1.50057	-0.11006	-0.46757	-0.42402	-0.04355
		1.41778	1.53335	-0.11557	-0.53068	-0.48712	-0.04356
		1.43109	1.54282	-0.11173	-0.48644	-0.44286	-0.04357
	M_2	1.42056	1.54656	-0.12600	-0.59028	-0.55501	-0.03527
		1.45213	1.58170	-0.12958	-0.48808	-0.45839	-0.02969
		1.44529	1.58427	-0.13898	-0.53270	-0.50204	-0.03067
	M_3	1.35978	1.49052	-0.13074	-0.51757	-0.47335	-0.04422
		1.45076	1.59029	-0.13952	-0.44537	-0.41521	-0.03016
		1.39188	1.50636	-0.11449	-0.44887	-0.41701	-0.03186
B	M_0	1.37812	1.52862	-0.15050	-0.42787	-0.40734	-0.02053
		1.33930	1.48954	-0.15024	-0.45999	-0.42705	-0.03294
		1.31131	1.46323	-0.15192	-0.50080	-0.47691	-0.02389
	M_1	1.37183	1.52055	-0.14872	-0.38971	-0.36474	-0.02496
		1.41628	1.56084	-0.14455	-0.35782	-0.33604	-0.02178
		1.41694	1.58145	-0.16452	-0.30525	-0.28627	-0.01897
	M_2	1.41335	1.53599	-0.12264	-0.36283	-0.33398	-0.02885
		1.38062	1.52951	-0.14889	-0.36526	-0.34397	-0.02129
		1.46776	1.60944	-0.14168	-0.30636	-0.27971	-0.02664
	M_3	1.31687	1.46183	-0.14497	-0.39744	-0.36277	-0.03467
		1.30043	1.44261	-0.14218	-0.46861	-0.43562	-0.03300
		1.30285	1.43305	-0.13020	-0.49519	-0.45605	-0.03913

Table 4.93 Physics B Average MC Item Parameter Estimates

		Item Parameter Estimates								
		a			b			c		
Form	Match	MC+FR	MC	Diff	MC+FR	MC	Diff	MC+FR	MC	Diff
A	M_0	0.95087	0.94652	0.00435	0.17913	0.16804	0.01109	0.13196	0.12565	0.00630
		1.00370	0.97565	0.02804	0.26022	0.22000	0.04022	0.14717	0.13065	0.01652
		0.98043	0.95543	0.02500	0.24870	0.22783	0.02087	0.13370	0.12500	0.00870
	M_1	0.93261	0.91022	0.02239	0.15783	0.13696	0.02087	0.12304	0.11283	0.01022
		0.99283	0.96630	0.02652	0.19826	0.16630	0.03196	0.14630	0.13022	0.01609
		0.93022	0.92826	0.00196	0.12783	0.10087	0.02696	0.14022	0.13391	0.00630
	M_2	0.93391	0.91000	0.02391	0.08870	0.06478	0.02391	0.12261	0.10783	0.01478
		0.97870	0.94217	0.03652	0.16152	0.09826	0.06326	0.13739	0.11522	0.02217
		0.91043	0.88783	0.02261	0.12261	0.07957	0.04304	0.13130	0.11413	0.01717
	M_3	0.95783	0.94348	0.01435	0.19500	0.18130	0.01370	0.12957	0.12239	0.00717
		0.95022	0.92283	0.02739	0.26022	0.22913	0.03109	0.13891	0.12217	0.01674
		0.94348	0.93239	0.01109	0.23652	0.22848	0.00804	0.13217	0.12935	0.00283
B	M_0	0.97043	0.93196	0.03848	0.07935	0.04130	0.03804	0.13913	0.11935	0.01978
		0.93043	0.90000	0.03043	-0.00674	-0.04696	0.04022	0.13174	0.11109	0.02065
		0.92065	0.89043	0.03022	-0.03739	-0.07435	0.03696	0.14304	0.12609	0.01696
	M_1	1.00043	0.95196	0.04848	0.21457	0.18304	0.03152	0.15717	0.14217	0.01500
		0.96152	0.90848	0.05304	0.14696	0.09500	0.05196	0.13739	0.11783	0.01957
		0.97891	0.94239	0.03652	0.11978	0.09152	0.02826	0.14609	0.13152	0.01457
	M_2	0.98152	0.93413	0.04739	0.16435	0.13196	0.03239	0.13761	0.12196	0.01565
		0.96739	0.91565	0.05174	0.14261	0.09087	0.05174	0.13261	0.10804	0.02457
		0.97957	0.93304	0.04652	0.09783	0.06239	0.03543	0.14848	0.13239	0.01609
	M_3	0.94174	0.89761	0.04413	0.09957	0.03283	0.06674	0.14022	0.10957	0.03065
		0.94717	0.89717	0.05000	0.04022	-0.01804	0.05826	0.13935	0.11217	0.02717
		0.92978	0.88848	0.04130	-0.04348	-0.08370	0.04022	0.13935	0.12239	0.01696

Table 4.94 Physics B Average FR Item Parameter Estimates

		Item Parameter Estimates					
		a			b		
Form	Match	MC+FR	FR	Diff	MC+FR	FR	Diff
A	M_0	1.96504	2.10222	-0.13718	0.46993	0.45988	0.01005
		1.92800	2.05551	-0.12752	0.52231	0.50478	0.01753
		1.96509	2.03926	-0.07417	0.53759	0.54343	-0.00584
	M_1	1.89412	2.03223	-0.13811	0.44321	0.43737	0.00584
		1.95874	2.07252	-0.11378	0.46145	0.44682	0.01463
		1.83126	1.95081	-0.11955	0.39240	0.37611	0.01629
	M_2	1.91508	2.00273	-0.08765	0.39036	0.38135	0.00900
		1.98706	2.09332	-0.10626	0.45490	0.44247	0.01243
		1.87438	1.97498	-0.10060	0.41910	0.40663	0.01247
	M_3	1.93780	2.03832	-0.10052	0.46915	0.45400	0.01515
		1.88222	1.99910	-0.11688	0.53413	0.52358	0.01055
		1.83051	1.93385	-0.10334	0.54117	0.52829	0.01288
B	M_0	1.55351	1.73707	-0.18356	0.52831	0.50939	0.01892
		1.61928	1.74039	-0.12111	0.43956	0.43067	0.00890
		1.57667	1.67325	-0.09657	0.41762	0.41244	0.00518
	M_1	1.61823	1.77861	-0.16038	0.64333	0.61560	0.02773
		1.55587	1.72922	-0.17336	0.65100	0.62441	0.02660
		1.55401	1.68091	-0.12690	0.58015	0.56837	0.01178
	M_2	1.70354	1.83407	-0.13053	0.62523	0.61279	0.01245
		1.60451	1.81530	-0.21079	0.63558	0.60921	0.02637
		1.59457	1.71541	-0.12084	0.54819	0.53761	0.01058
	M_3	1.65621	1.83322	-0.17701	0.50688	0.48504	0.02185
		1.61208	1.78379	-0.17171	0.48876	0.47297	0.01579
		1.62547	1.76568	-0.14021	0.38594	0.37799	0.00794



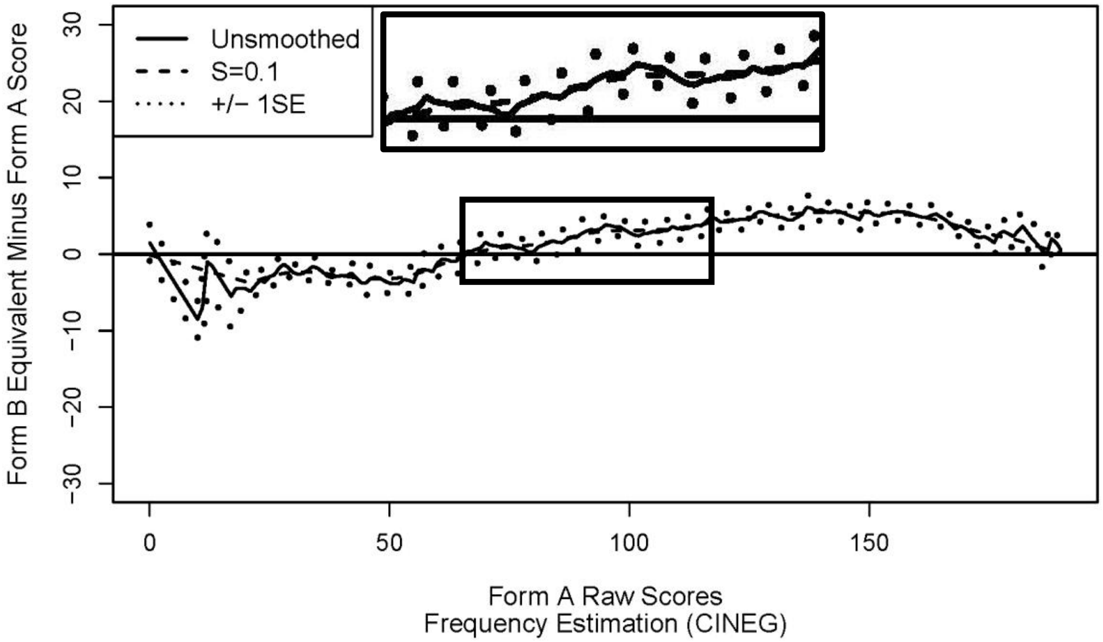


Figure 4.1 Comparison of unsmoothed and smoothed frequency estimation equivalents for Chemistry.

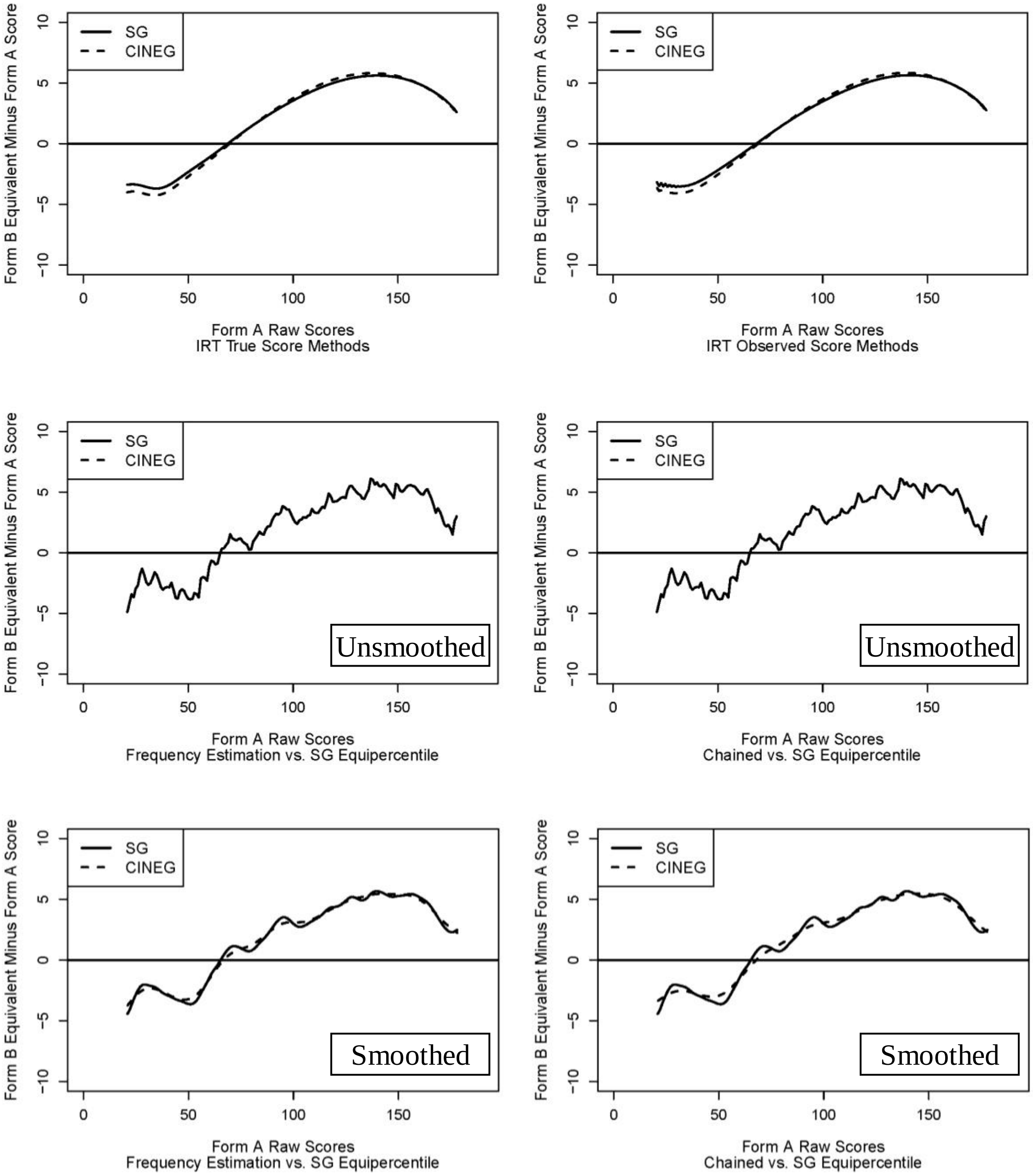


Figure 4.2 Comparison of SG and CINEG results for Chemistry.

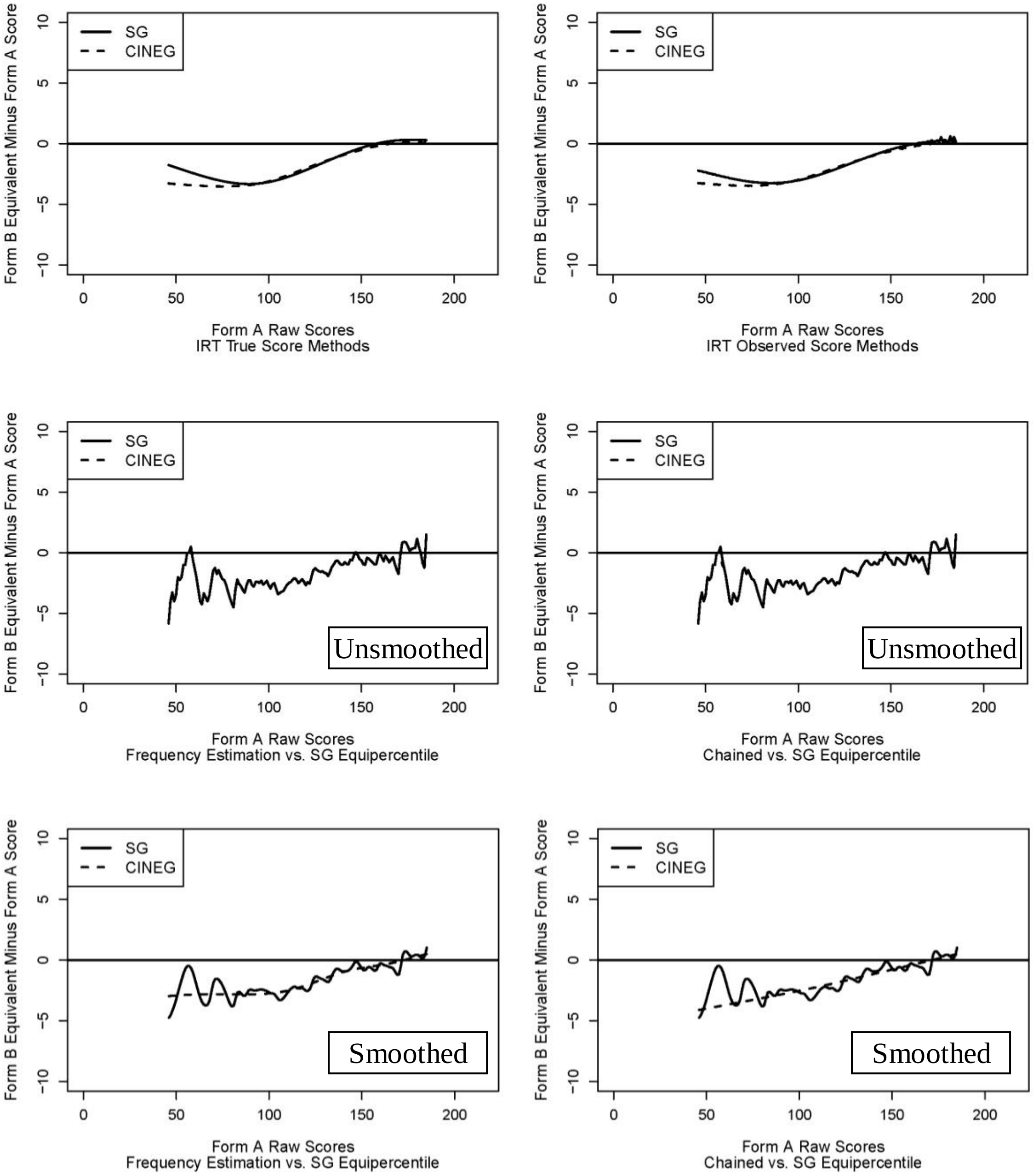


Figure 4.3 Comparison of SG and CINEG results for English Language.

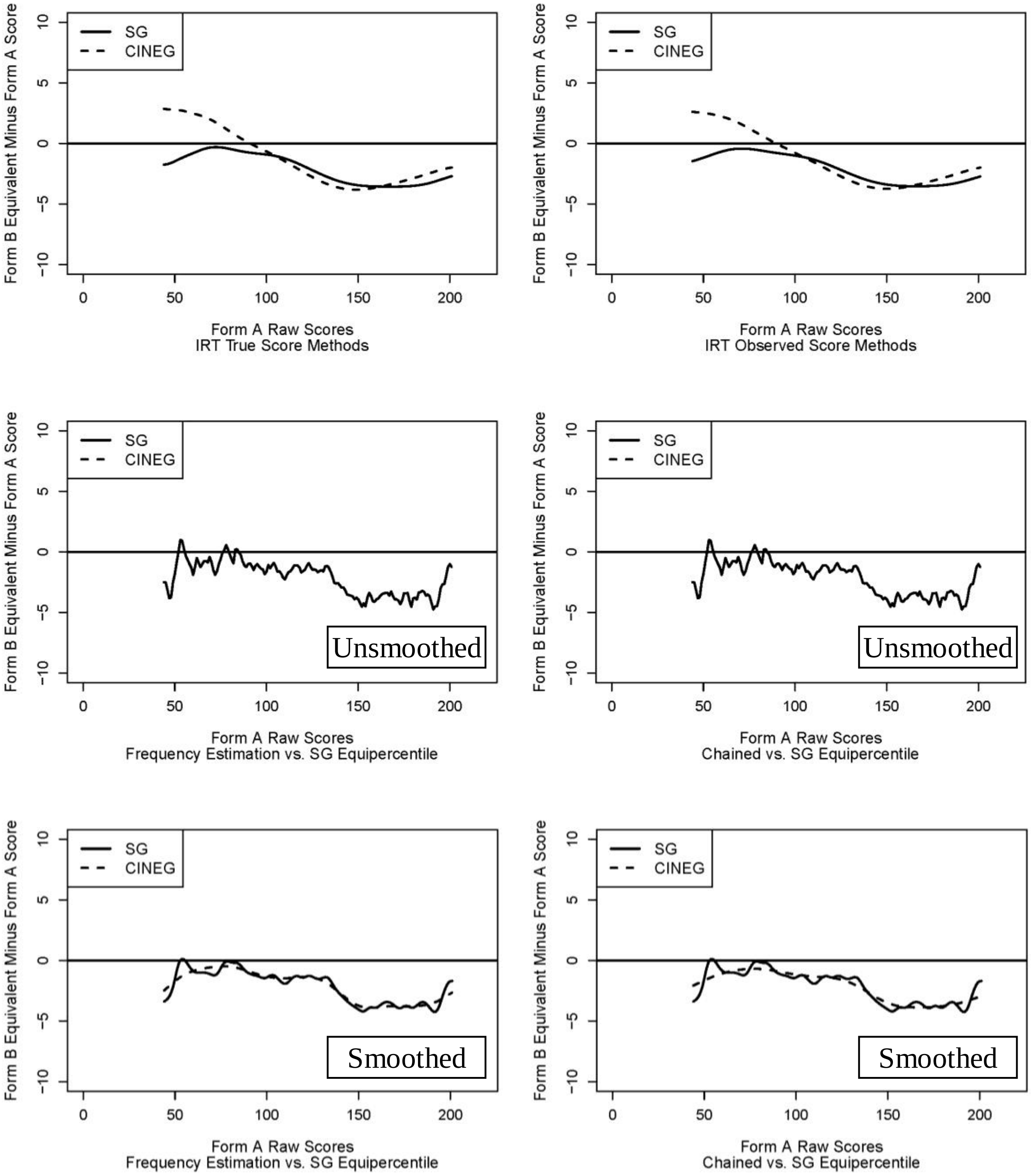


Figure 4.4 Comparison of SG and CINEG results for French Language.

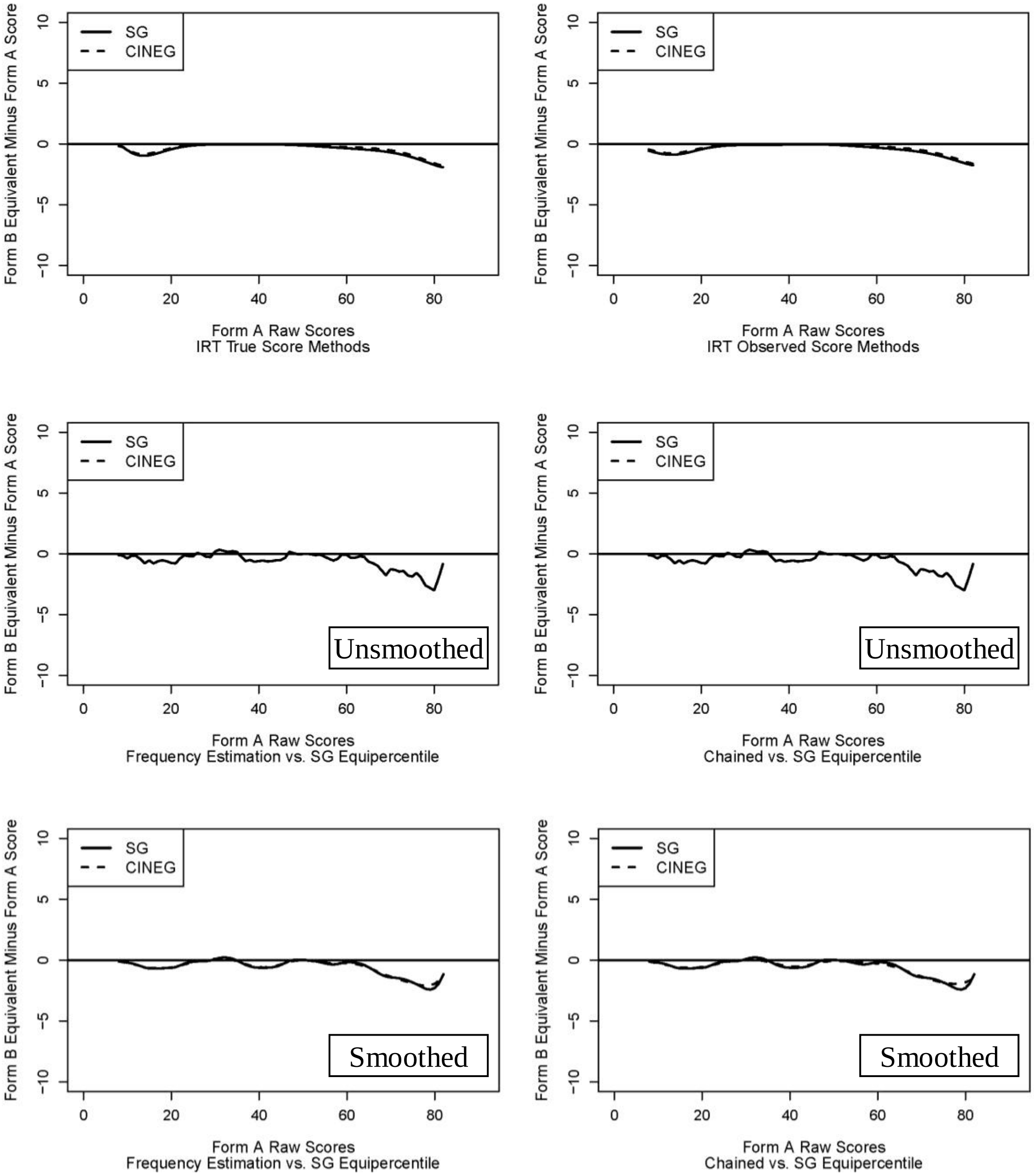


Figure 4.5 Comparison of SG and CINEG results for Physics B.

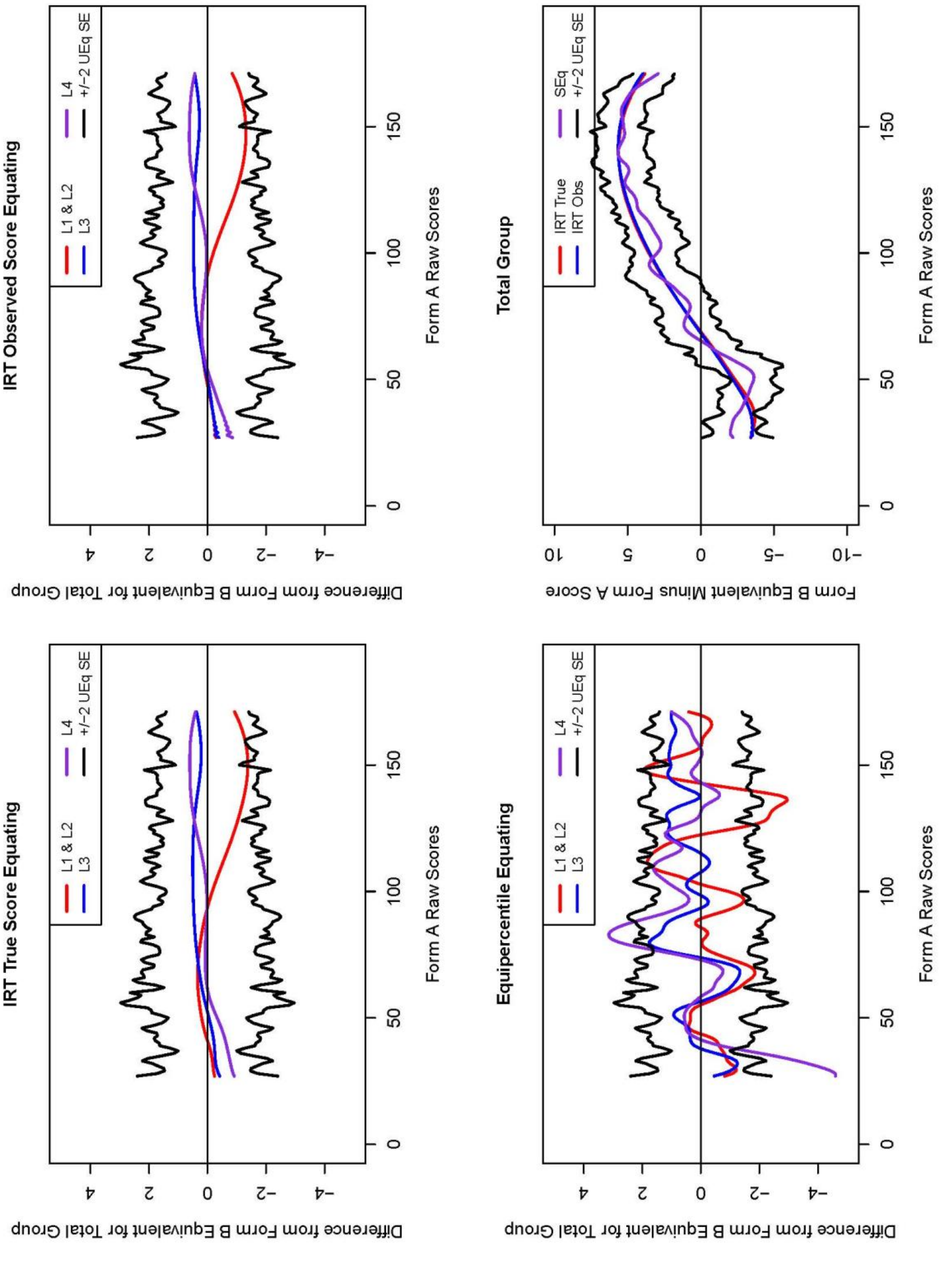


Figure 4.6 Comparison of SG equating relationships for Chemistry subgroups.

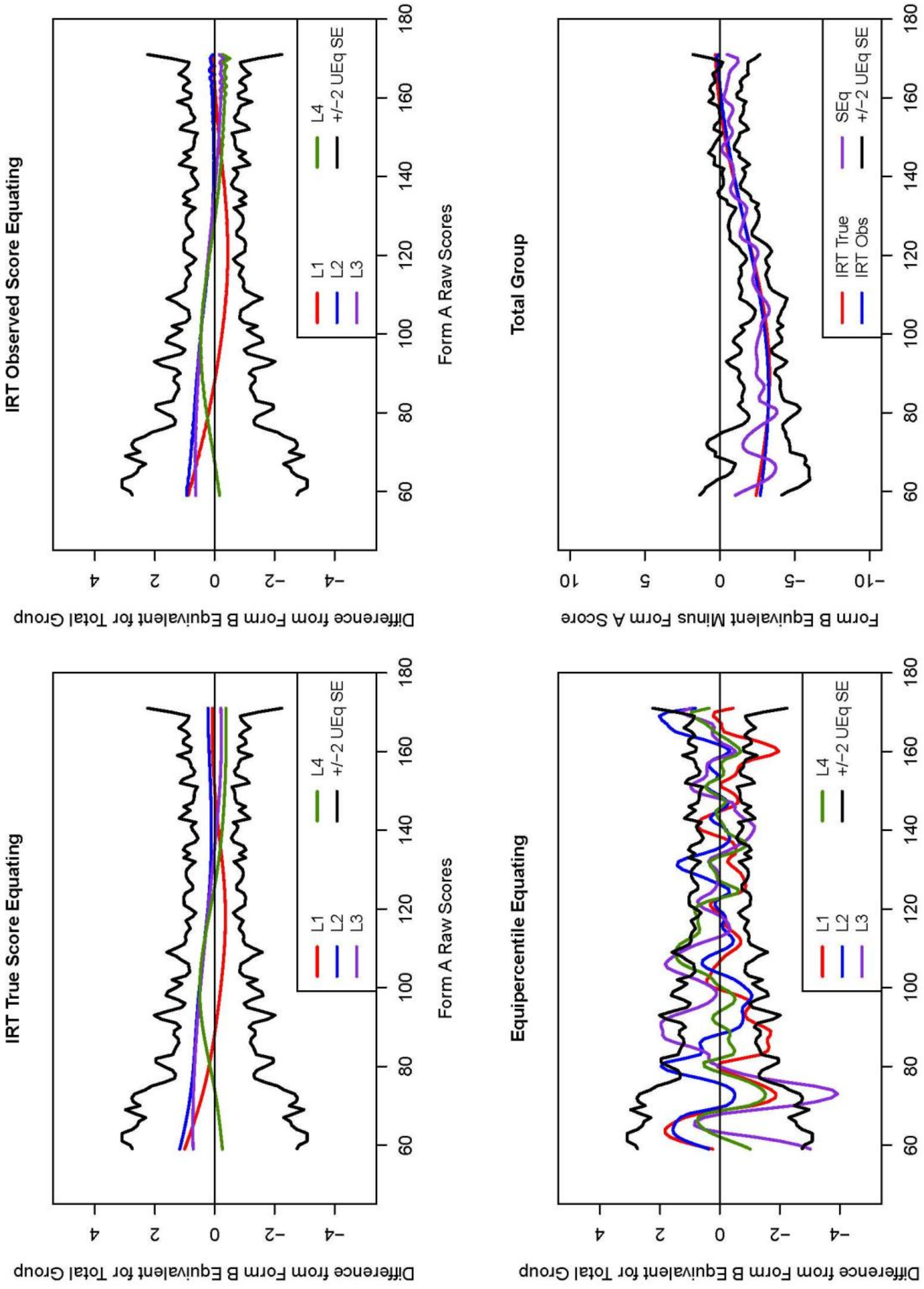


Figure 4.7 Comparison of SG equating relationships for English Language subgroups.

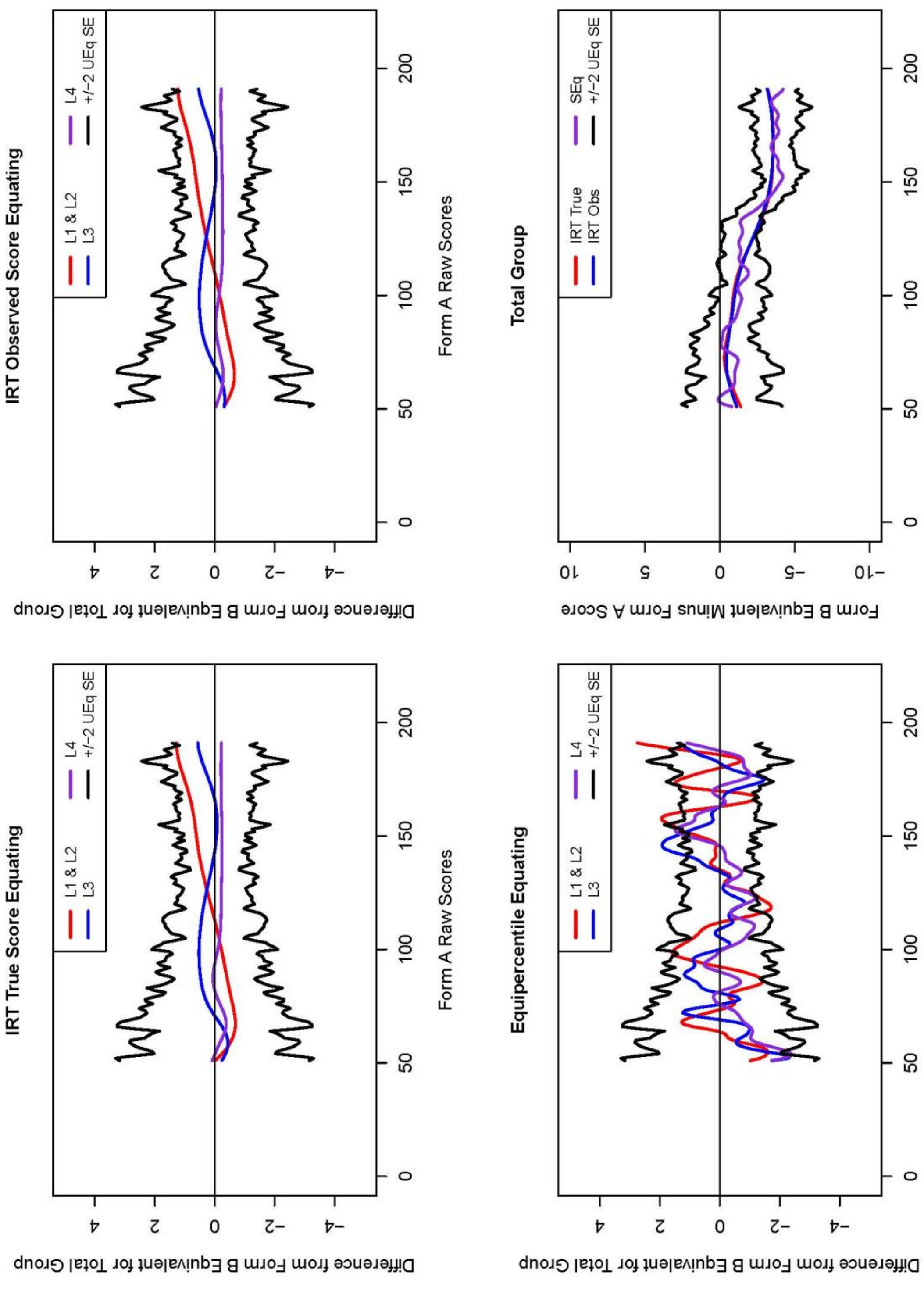


Figure 4.8 Comparison of SG equating relationships for French Language subgroups.

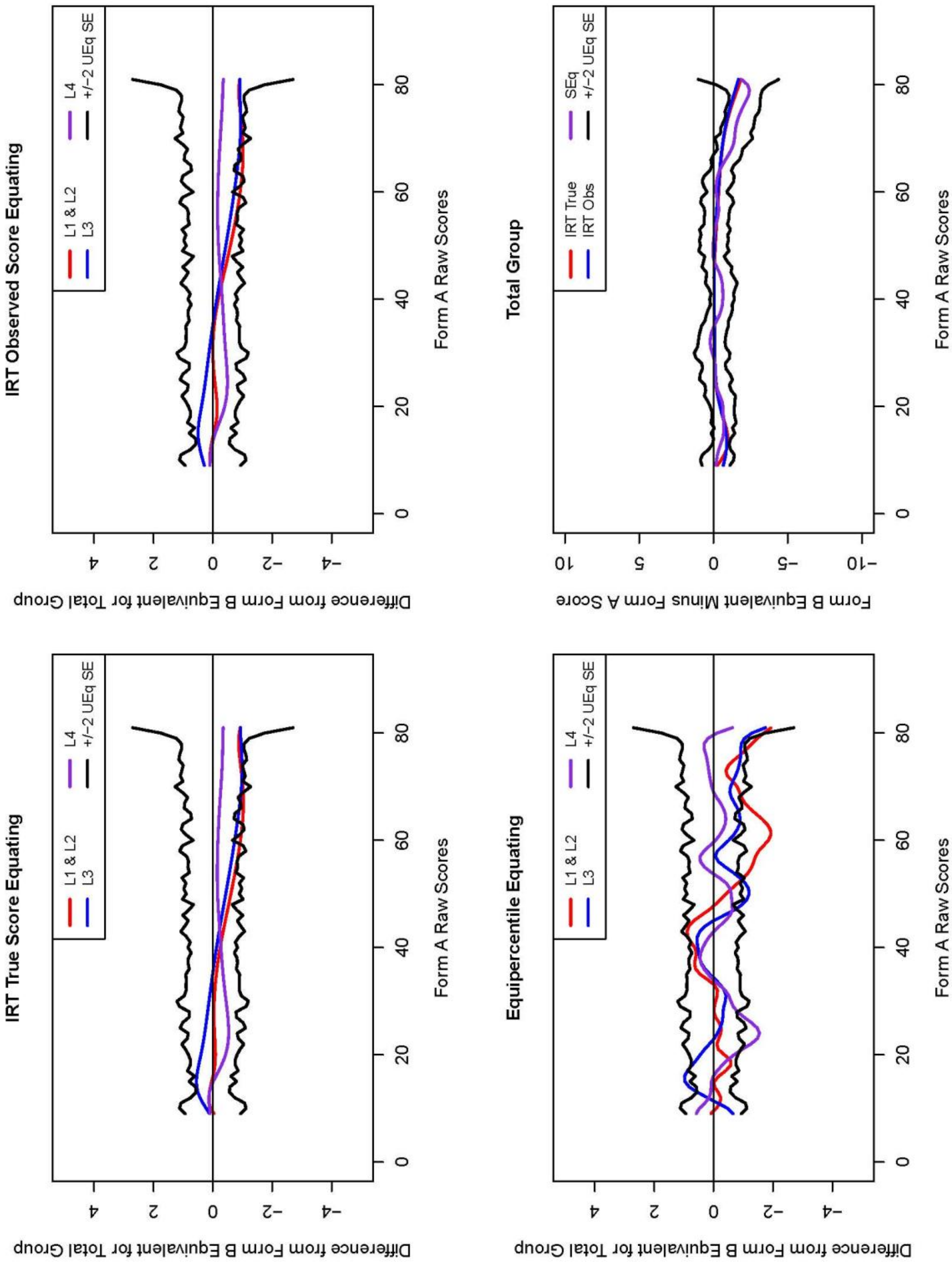


Figure 4.9 Comparison of SG equating relationships for Physics B subgroups.

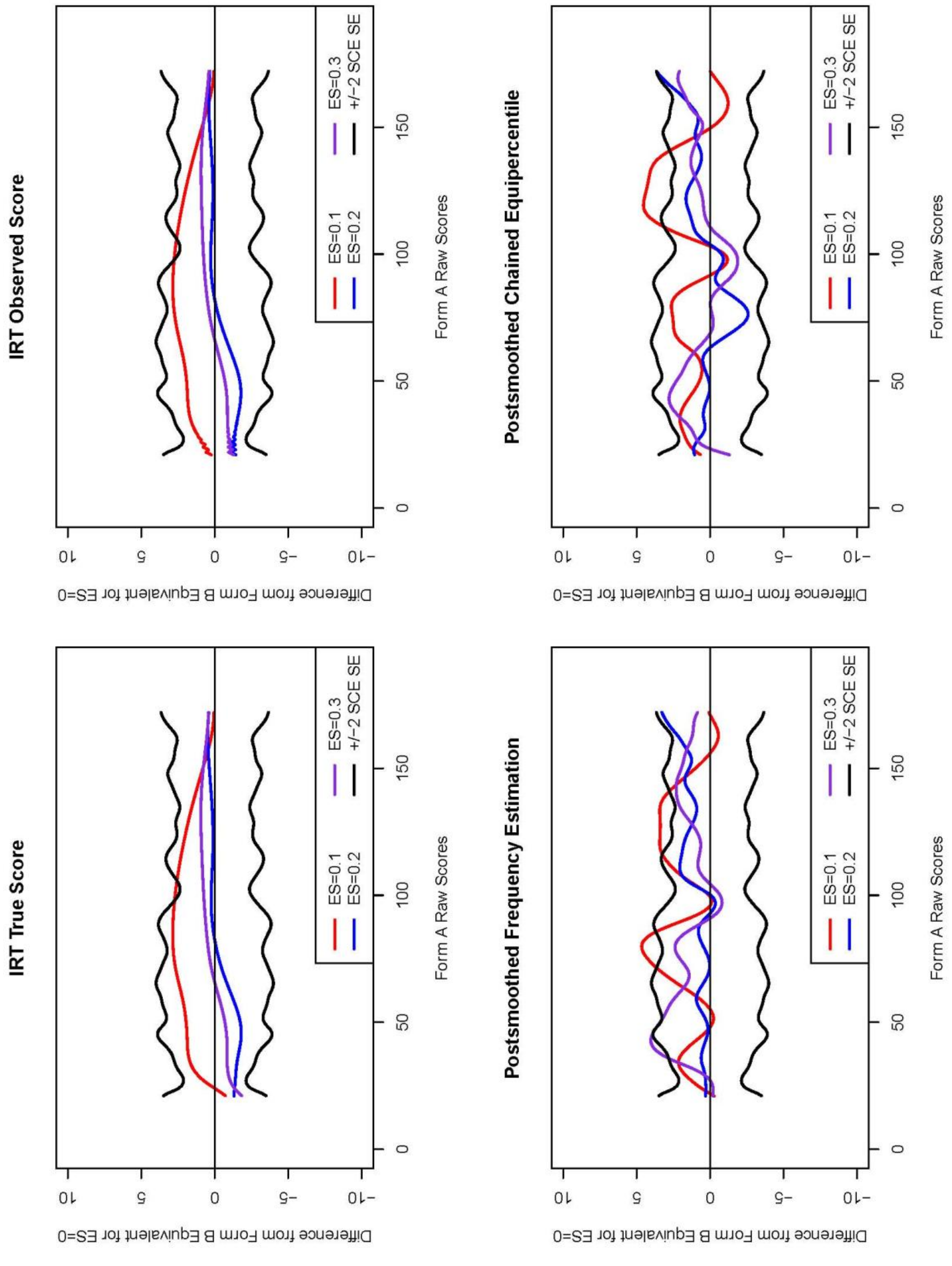


Figure 4.10 Comparison of Chemistry equating relationships for four ES levels.

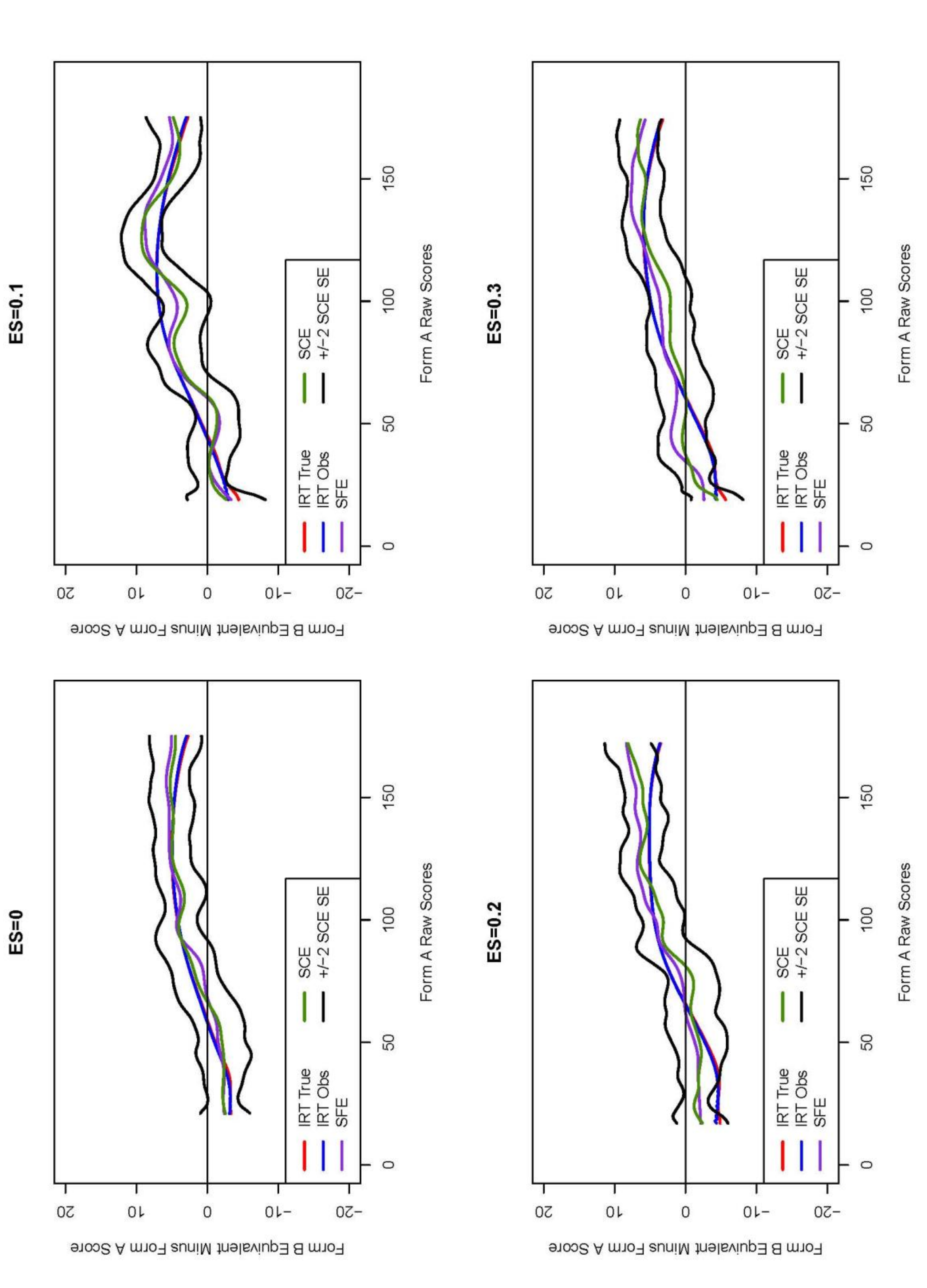


Figure 4.11 Comparison of Chemistry equating relationships for four equating methods.

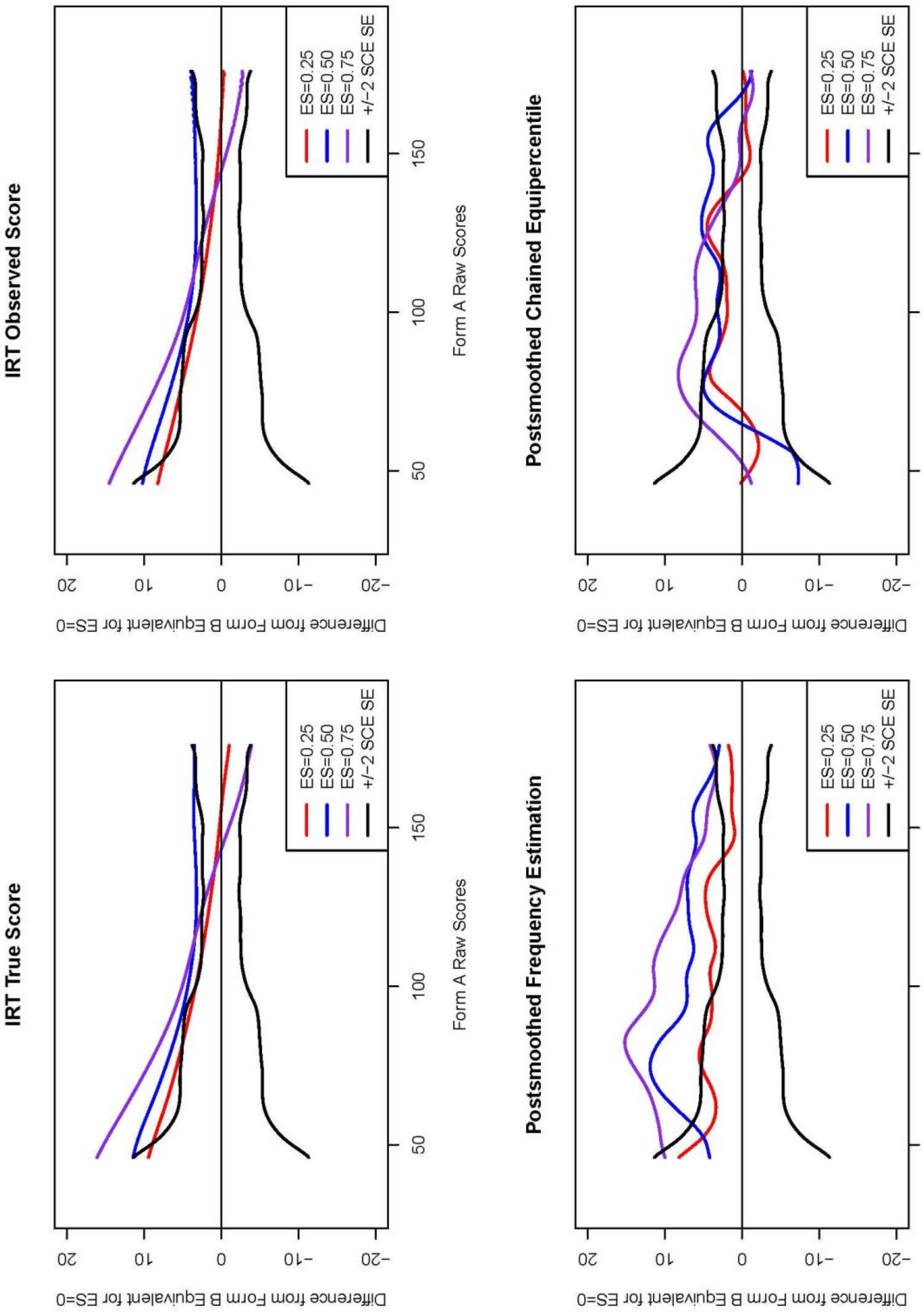


Figure 4.12 Comparison of English Language equating relationships for four ES levels.

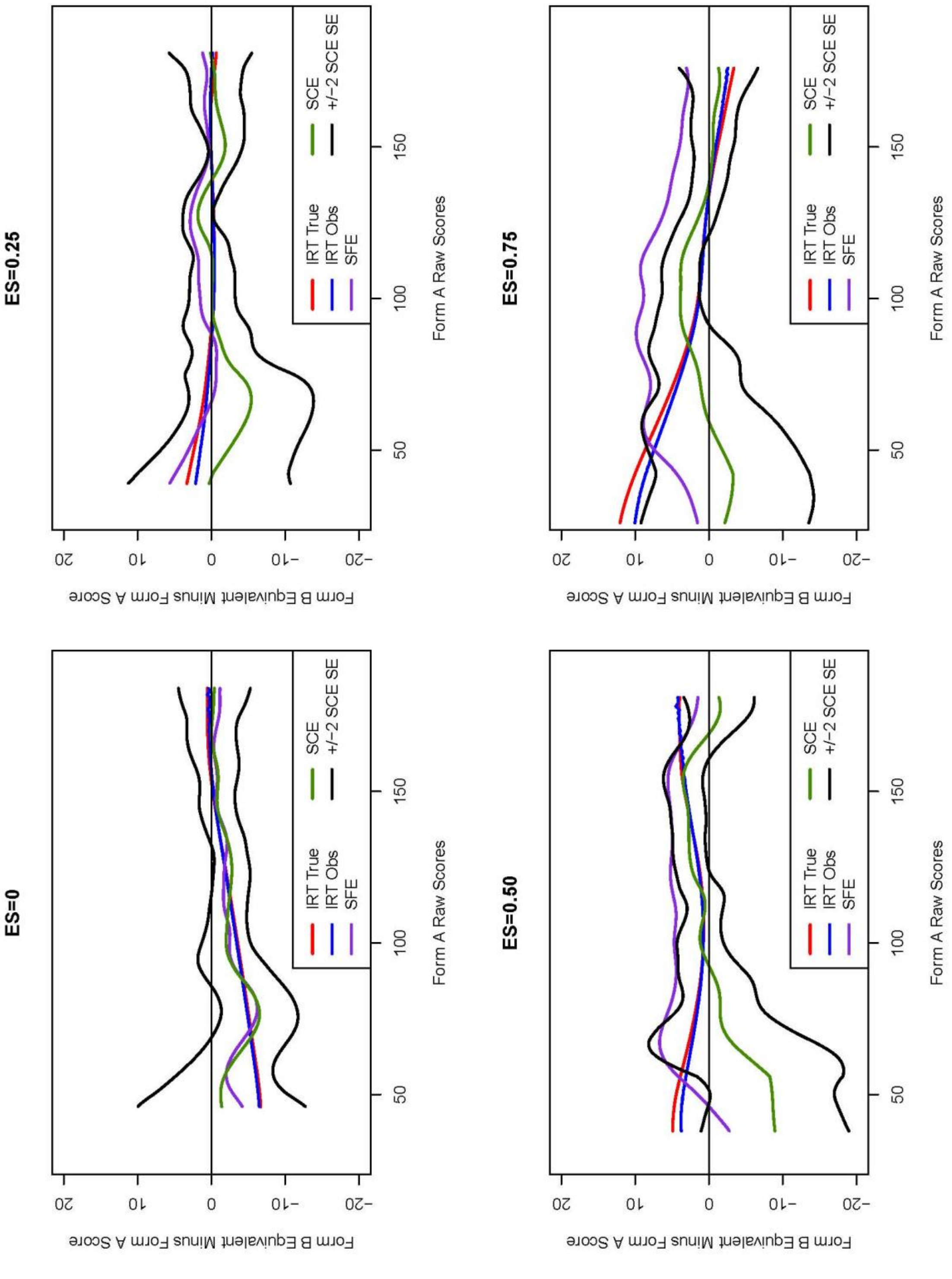


Figure 4.13 Comparison of English Language equating relationships for four equating methods.

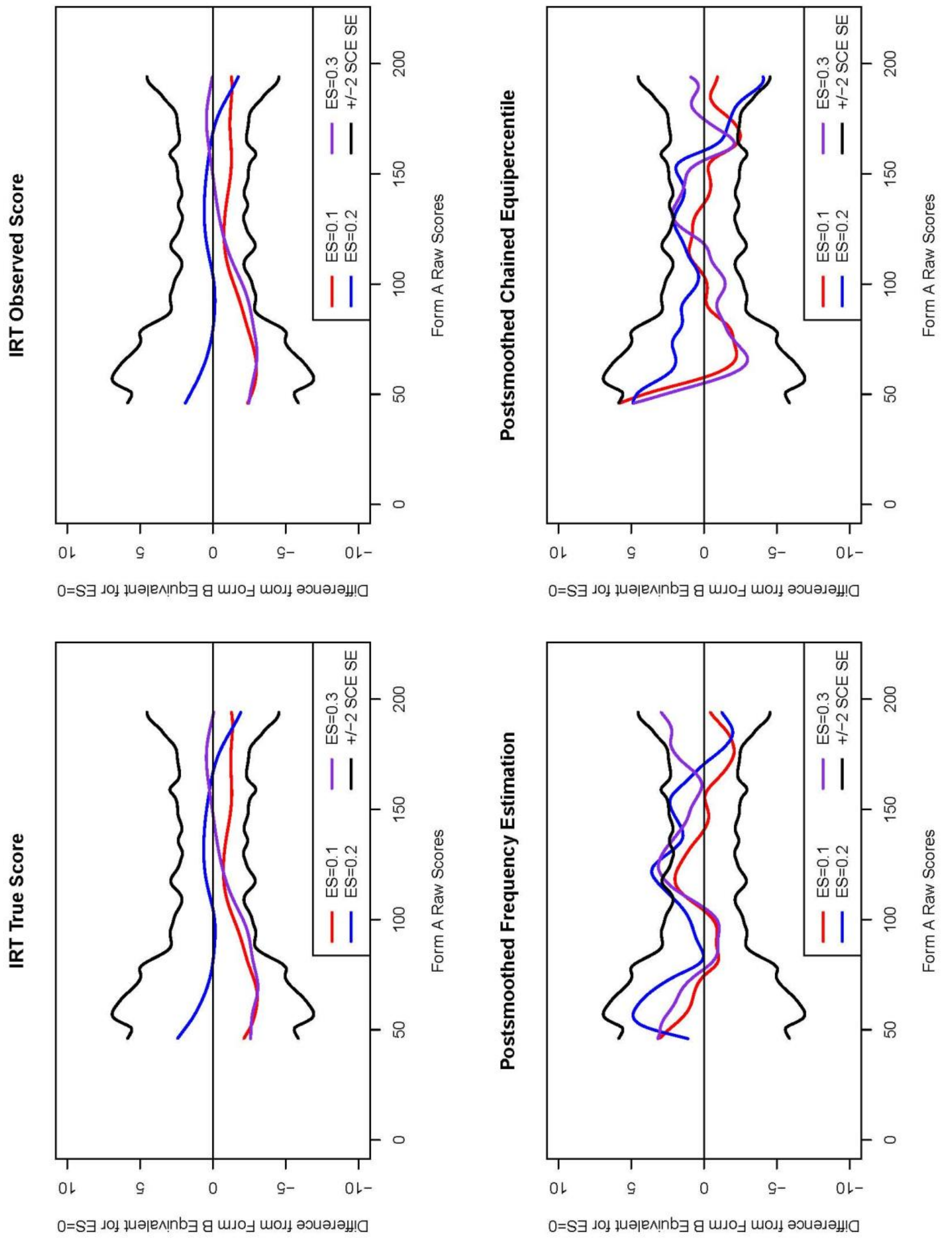


Figure 4.14 Comparison of French Language equating relationships for four ES levels.

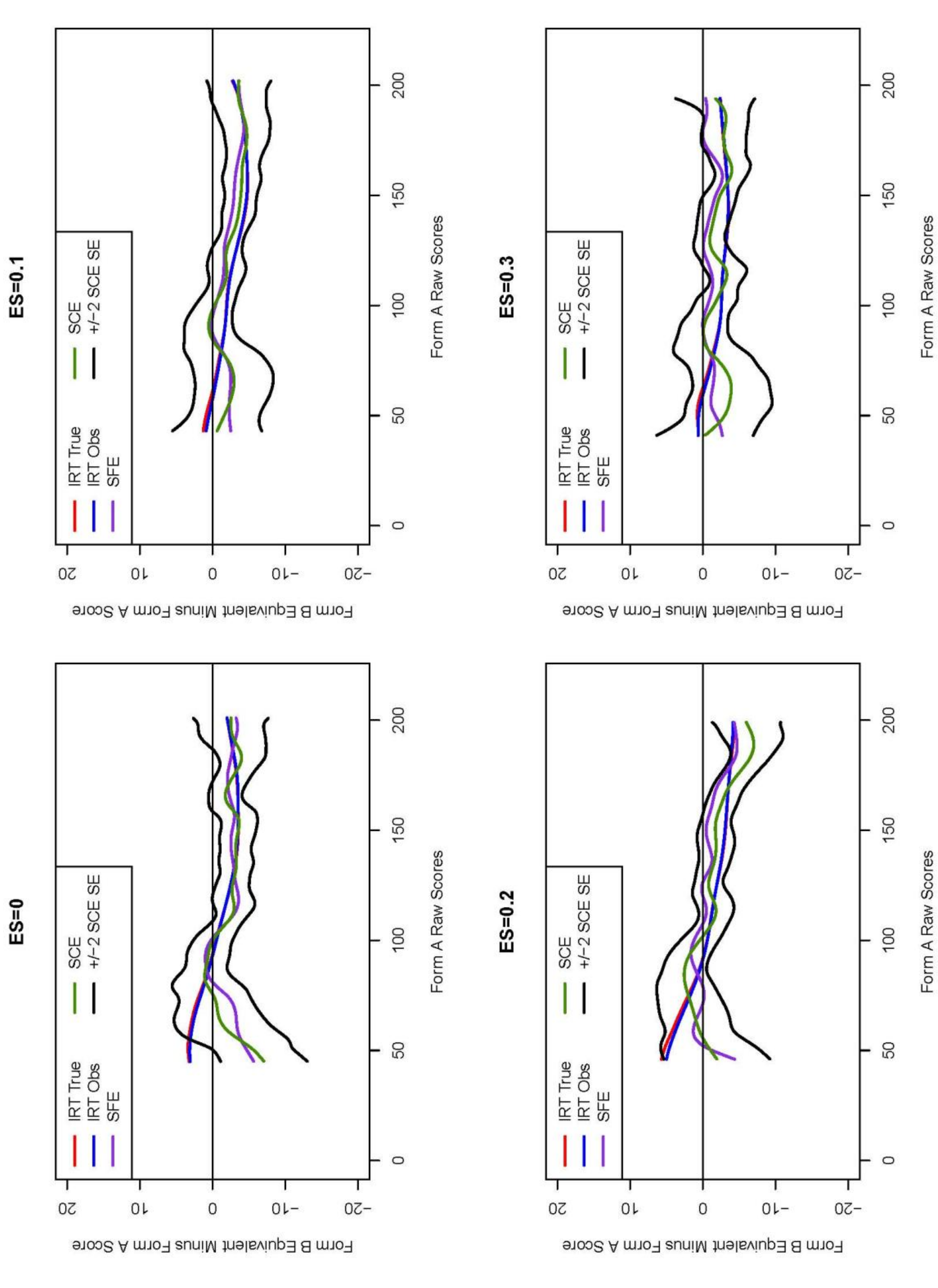


Figure 4.15 Comparison of French Language equating relationships for four equating methods.

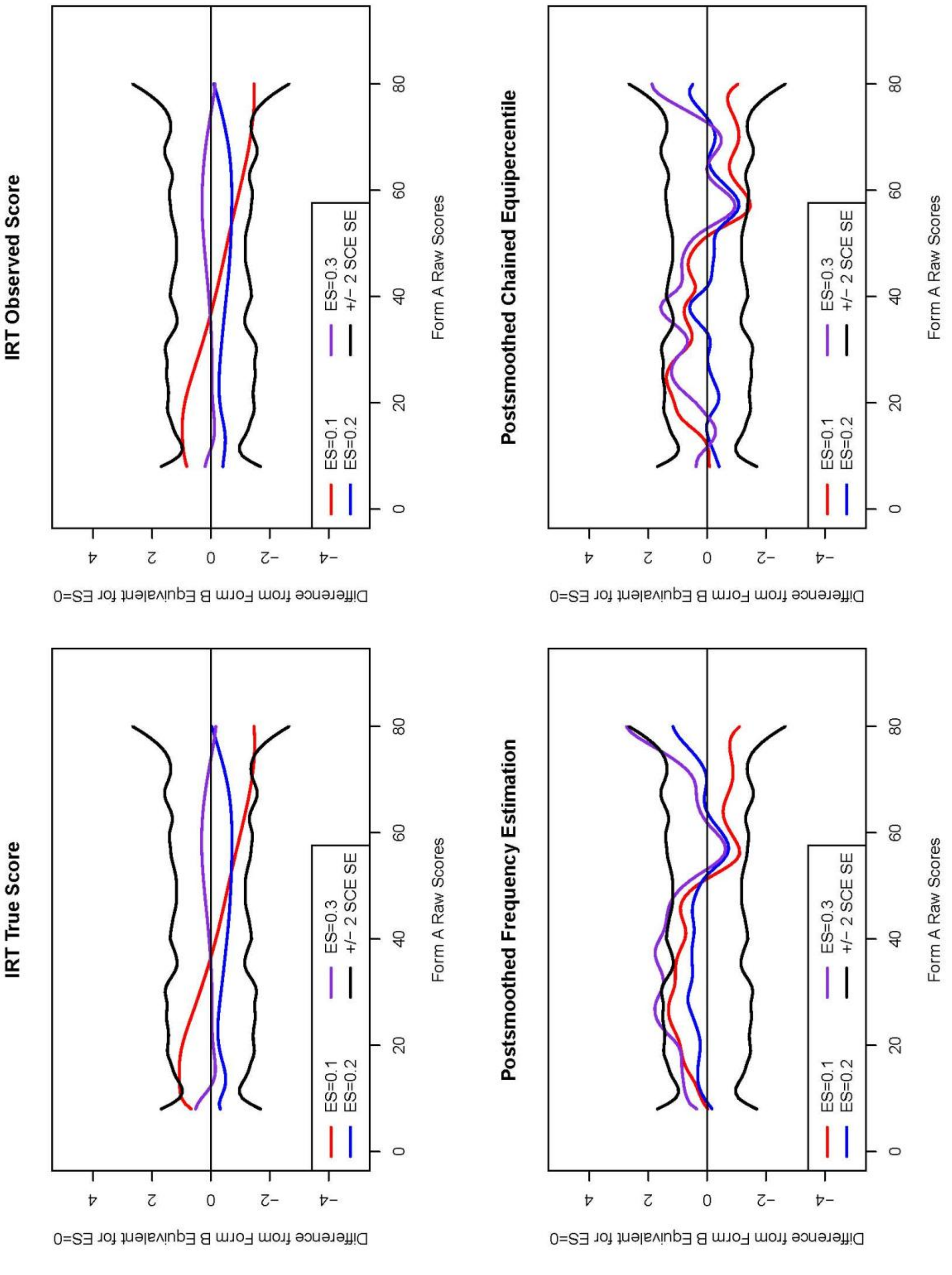


Figure 4.16 Comparison of Physics B equating relationships for four ES levels.

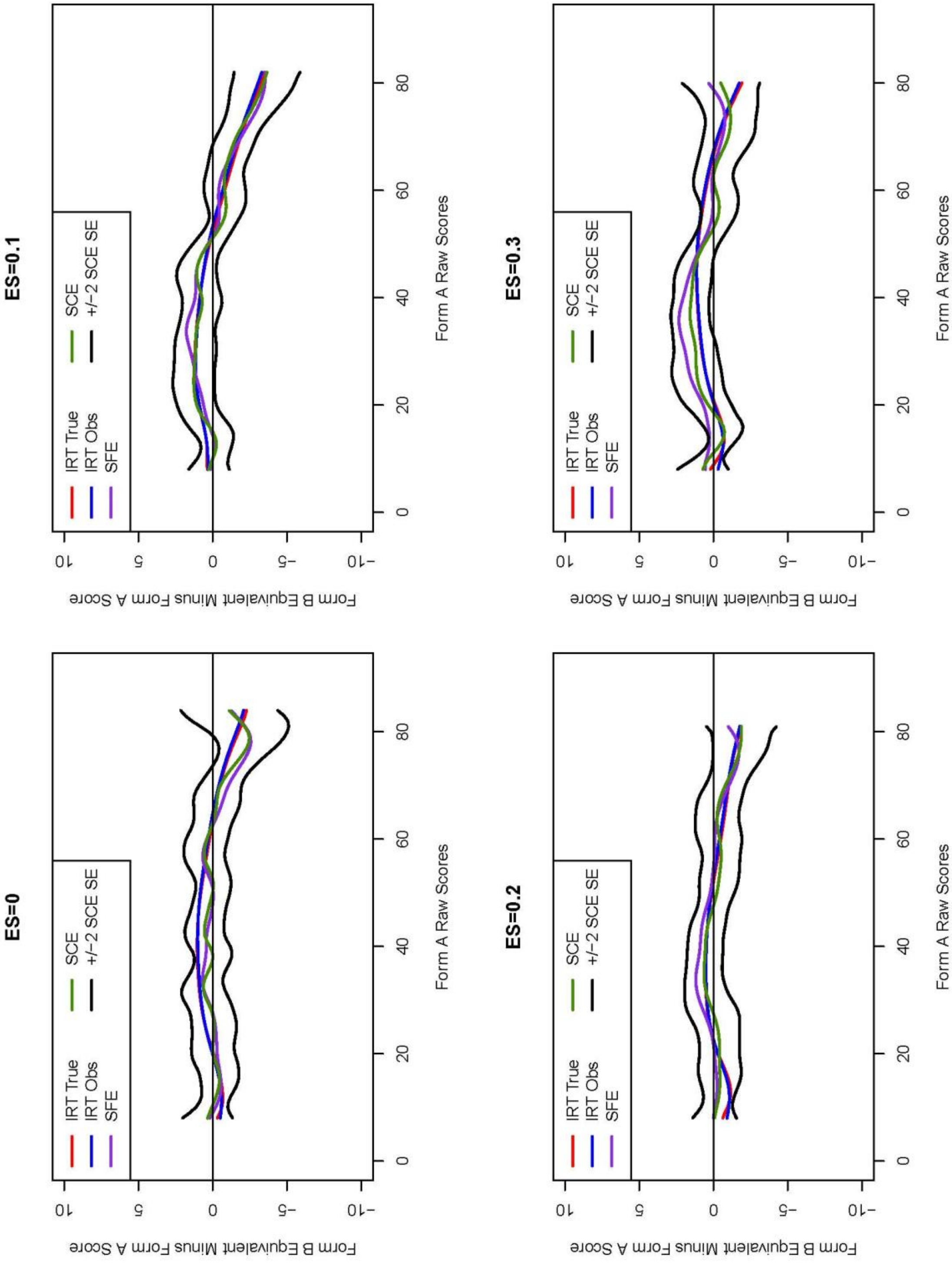
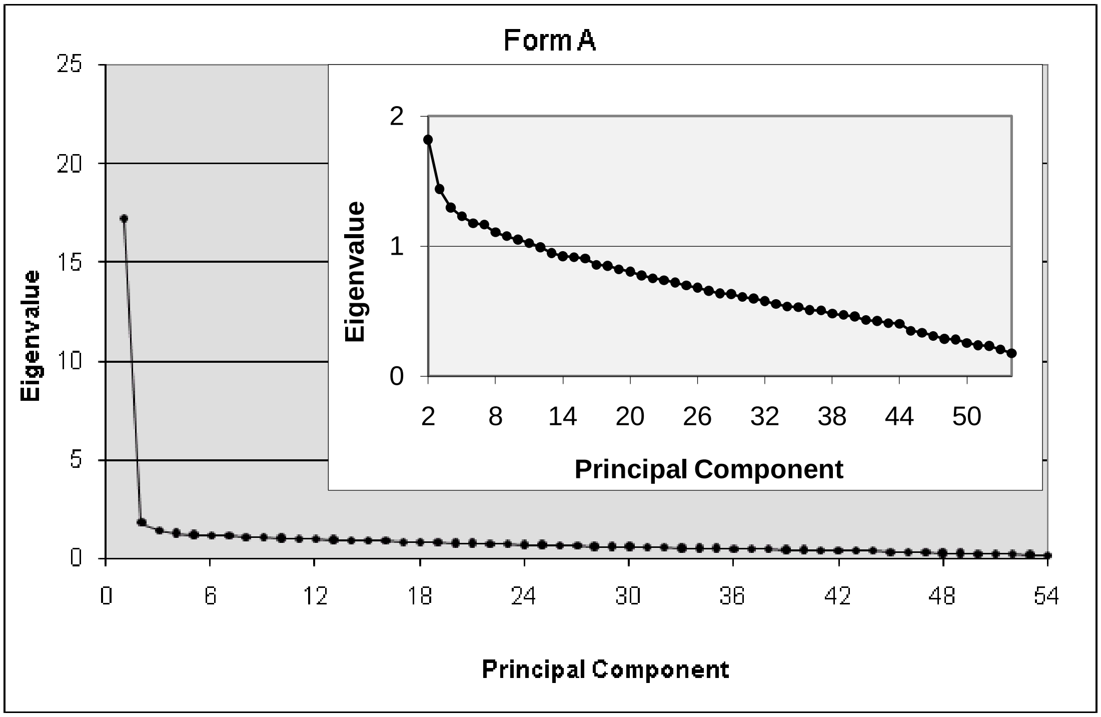


Figure 4.17 Comparison of Physics B equating relationships for four equating methods.



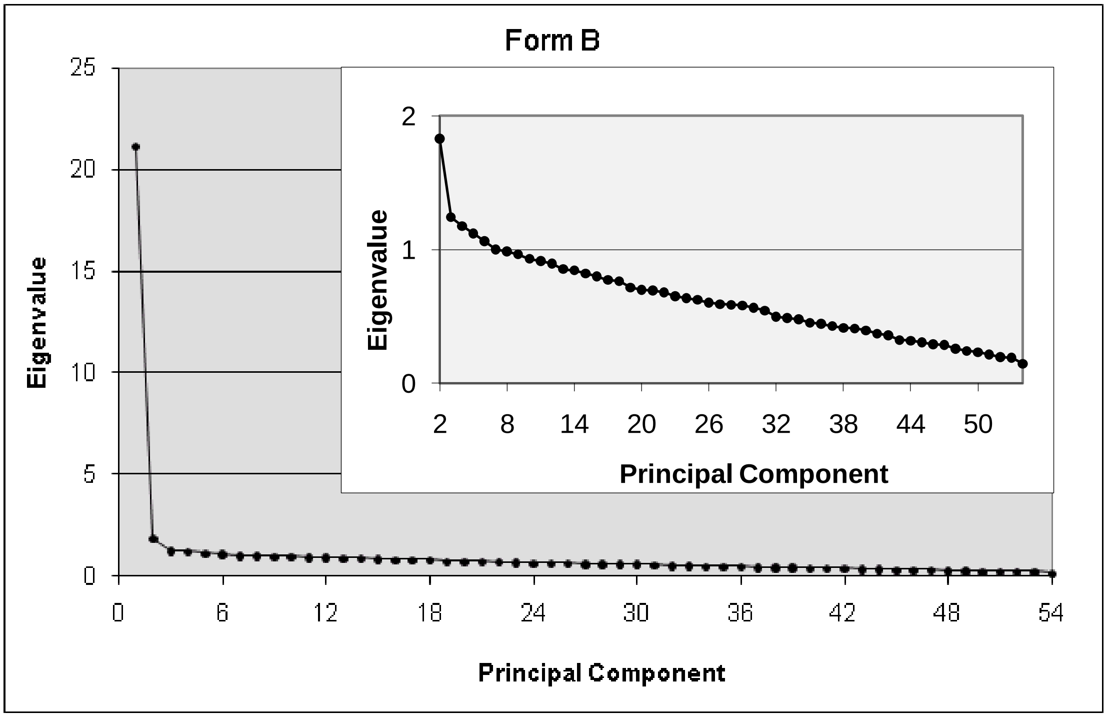
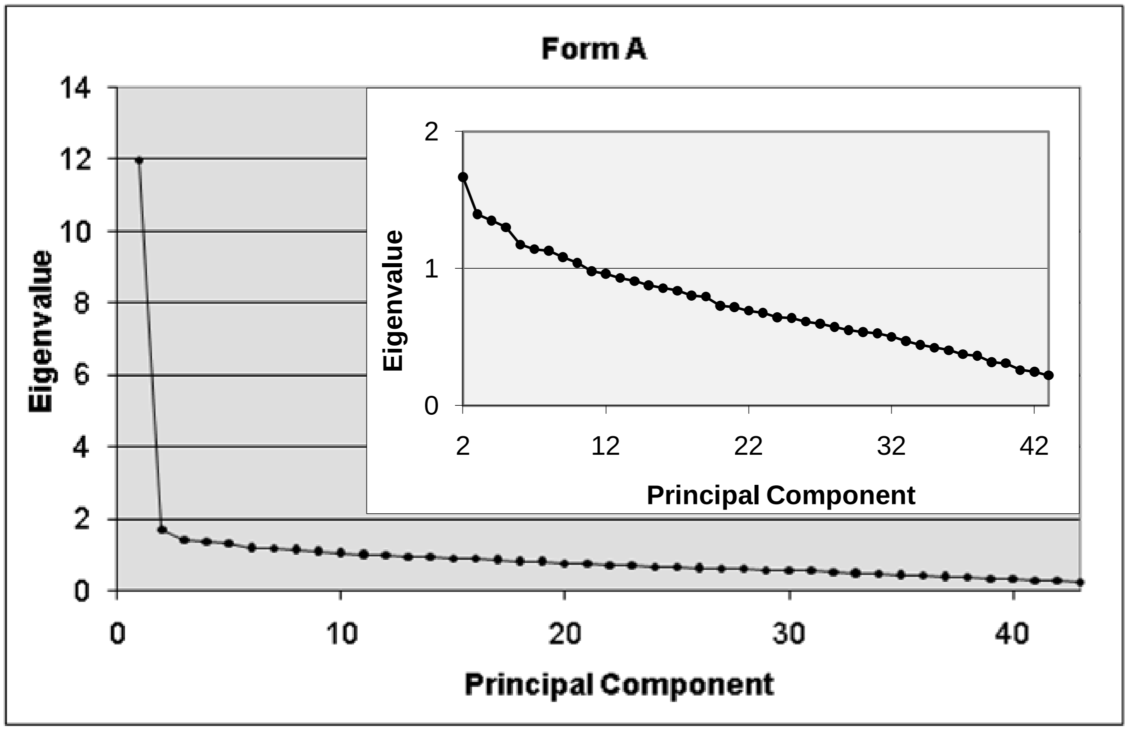


Figure 4.18 Scree Plots for Chemistry Form A and Form B where ES=0.



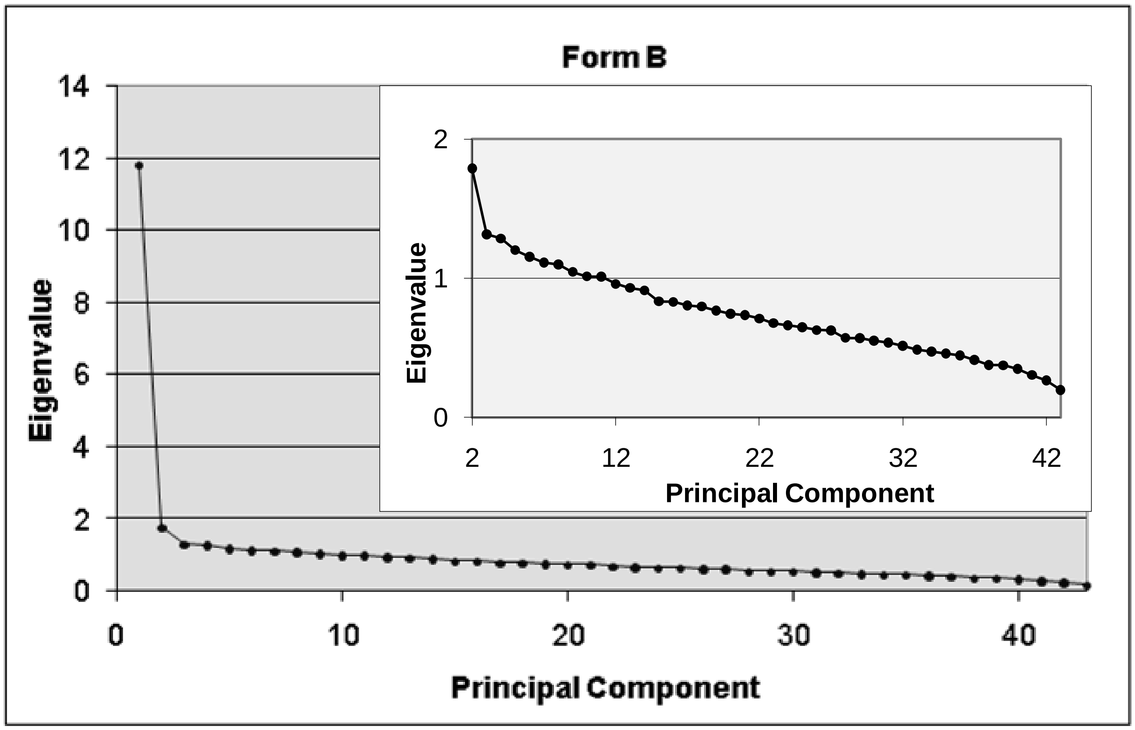


Figure 4.19 Scree Plots for English Language Form A and Form B where ES=0.

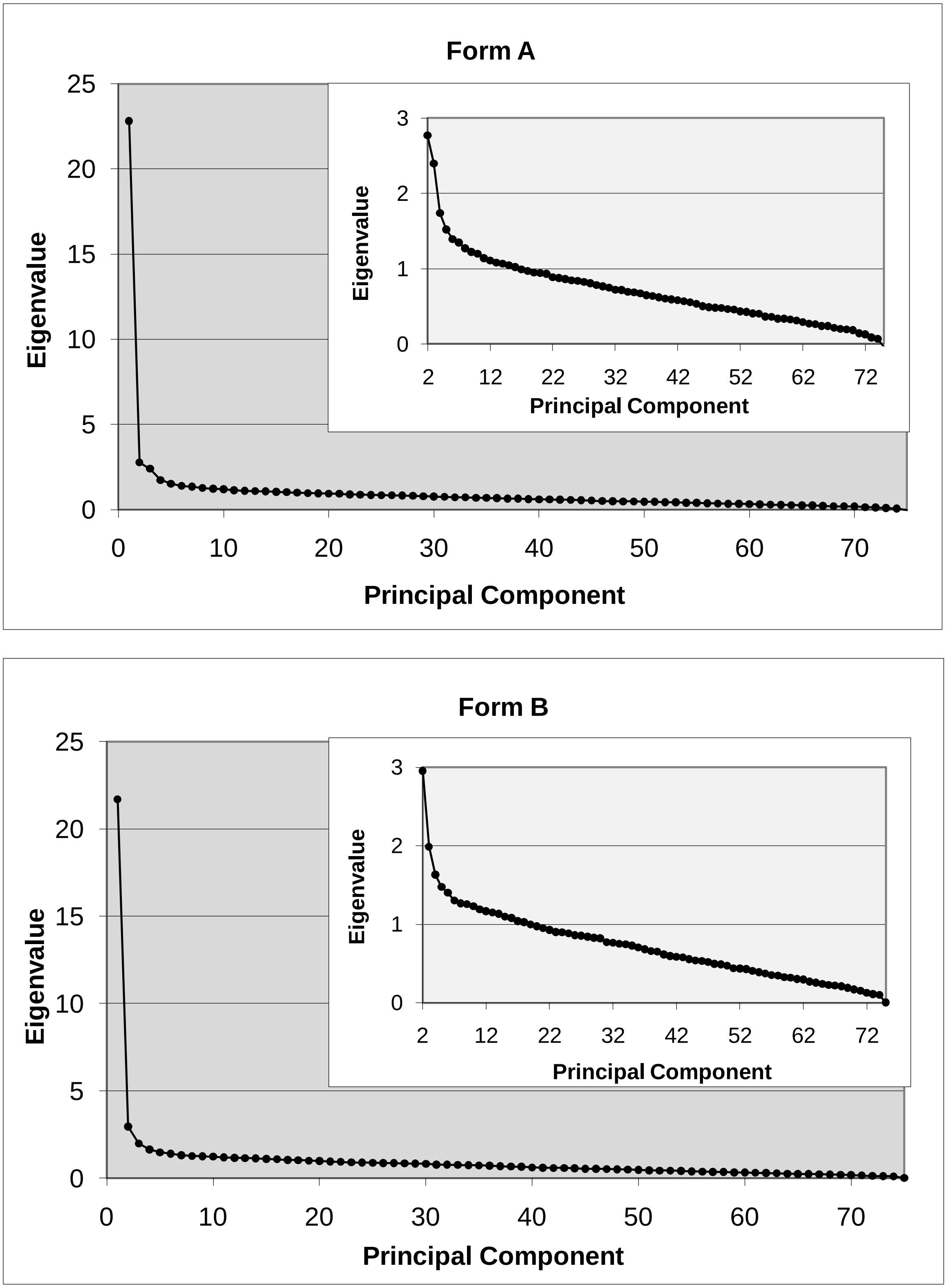


Figure 4.20 Scree Plots for French Language Form A and Form B.

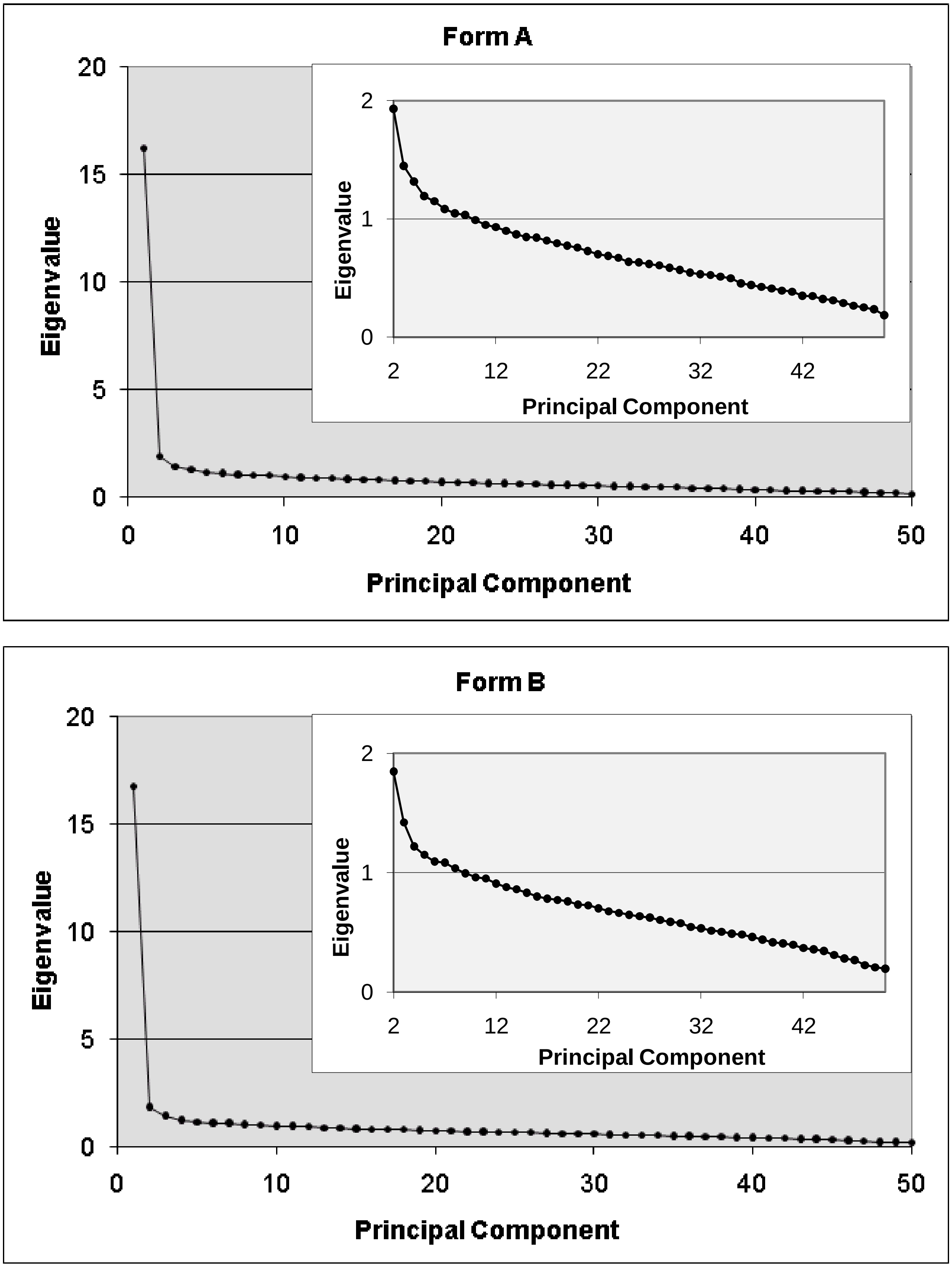


Figure 4.21 Scree Plots for Physics B Form A and Form B where ES=0.

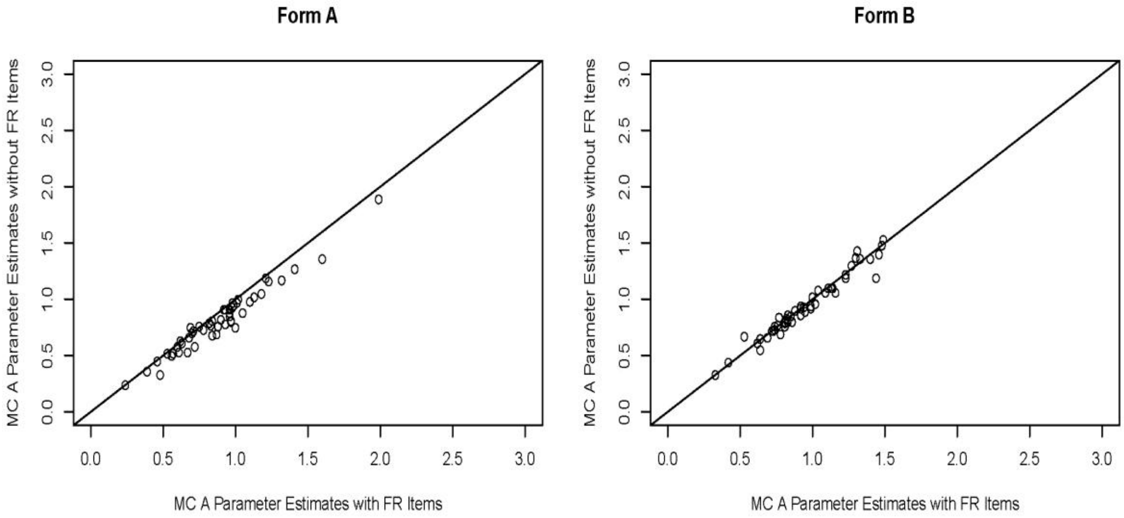


Figure 4.22 Comparison of Chemistry MC item *a*-parameter estimates for ES=0.

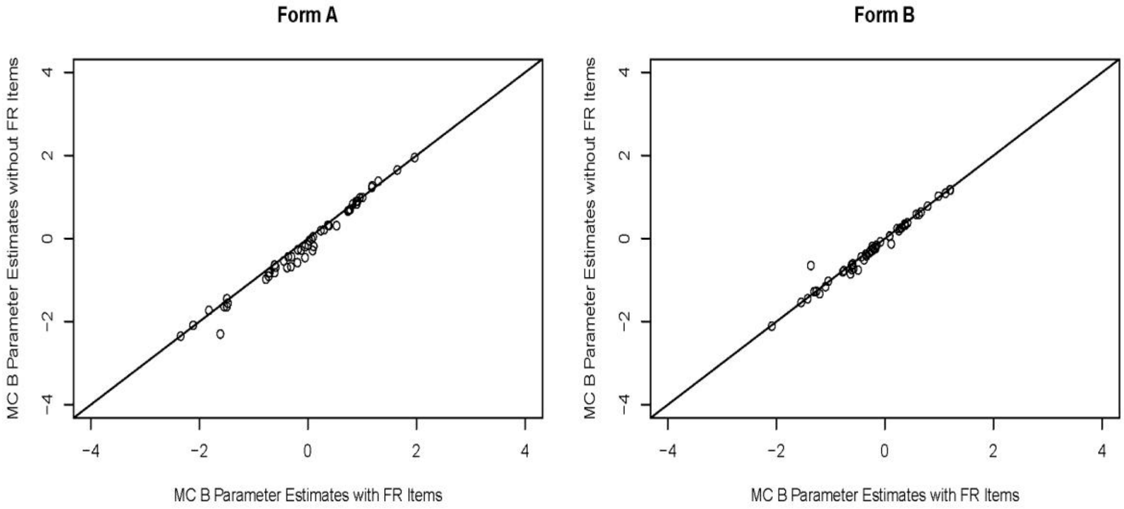


Figure 4.23 Comparison of Chemistry MC item *b*-parameter estimates for ES=0.

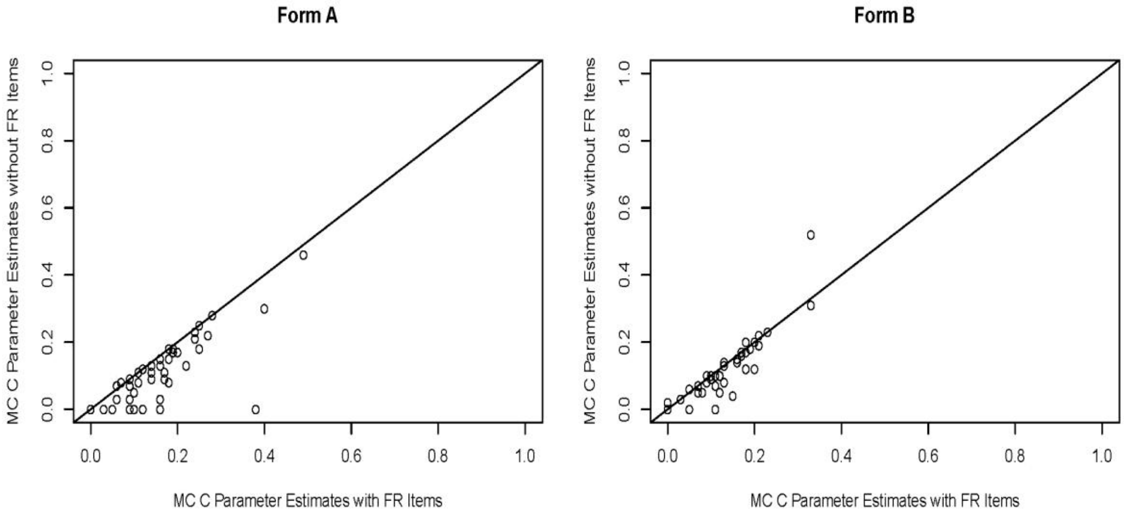


Figure 4.24 Comparison of Chemistry MC item *c*-parameter estimates for ES=0.

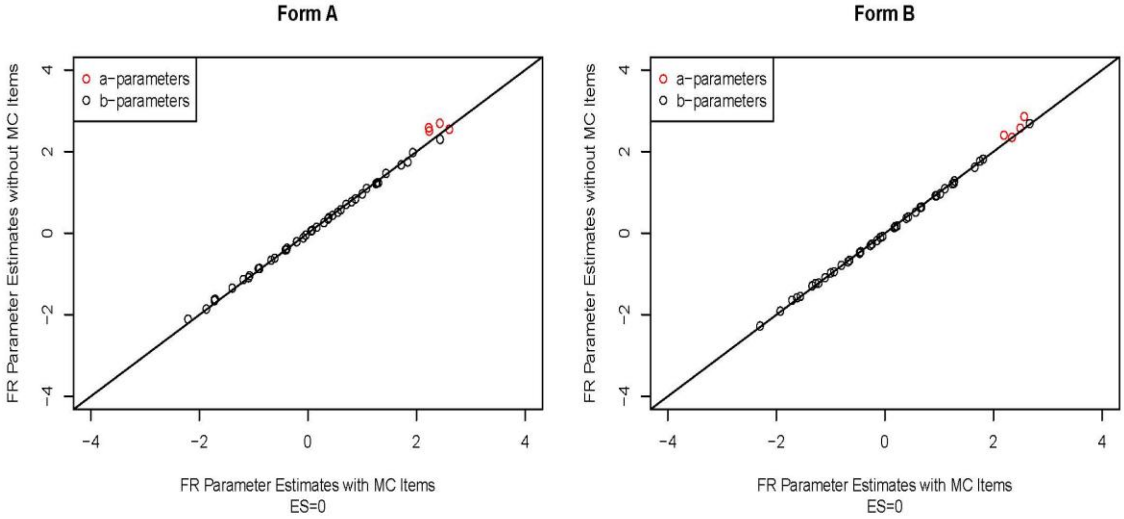


Figure 4.25 Comparison of Chemistry FR item parameter estimates for ES=0.

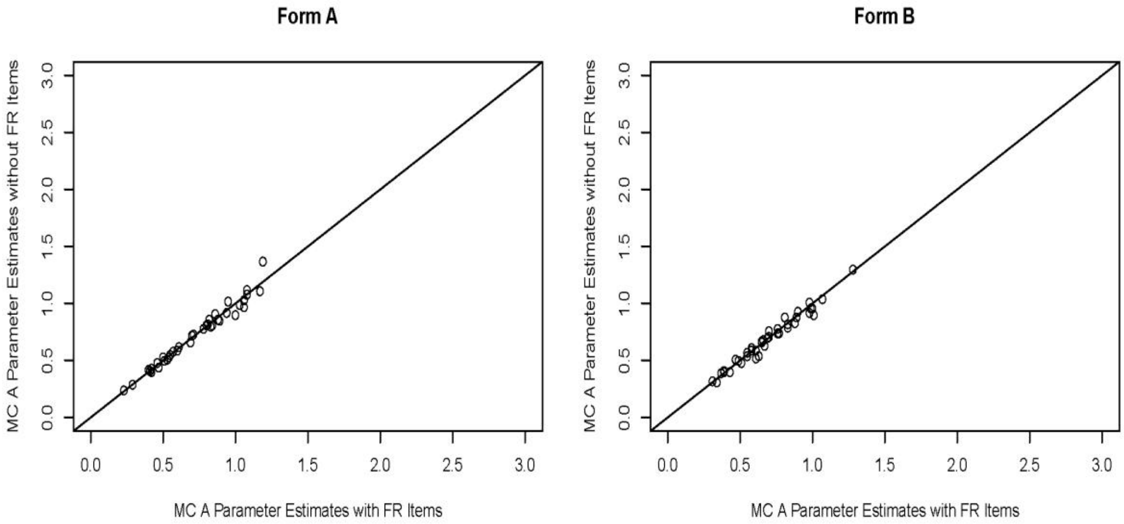


Figure 4.26 Comparison of English Language MC item *a*-parameter estimates for ES=0.

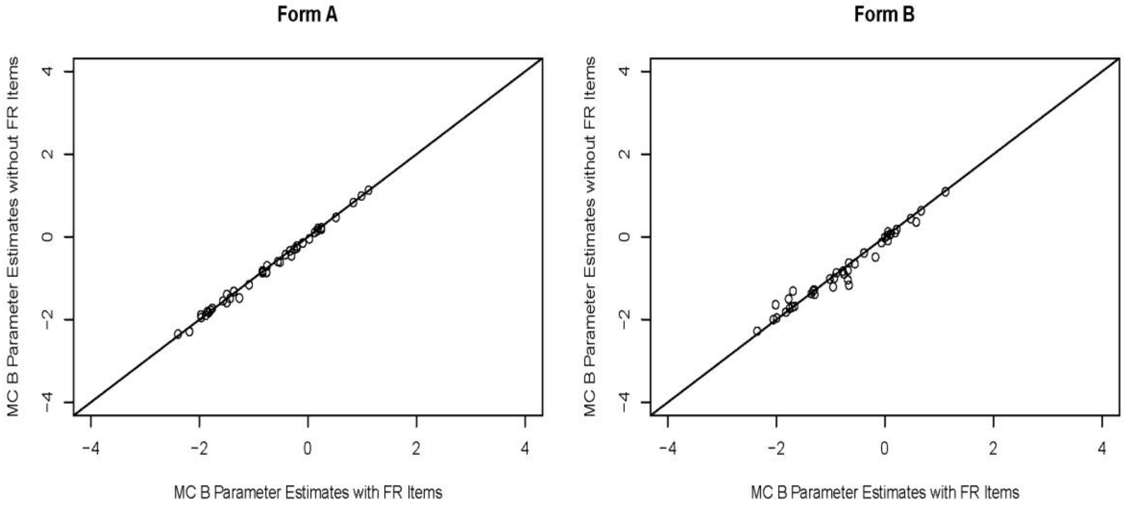


Figure 4.27 Comparison of English Language MC item *b*-parameter estimates for ES=0.

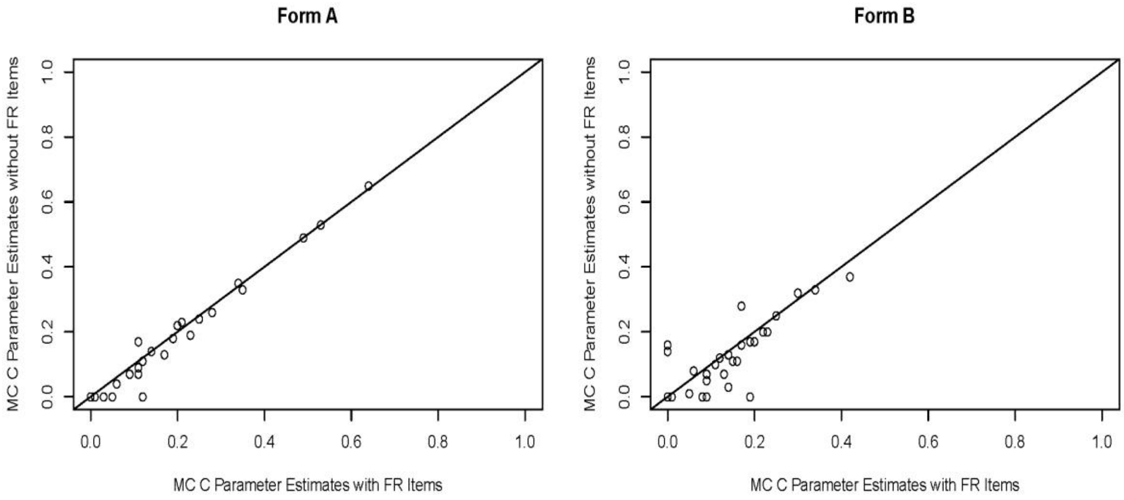


Figure 4.28 Comparison of English Language MC item *c*-parameter estimates for ES=0.

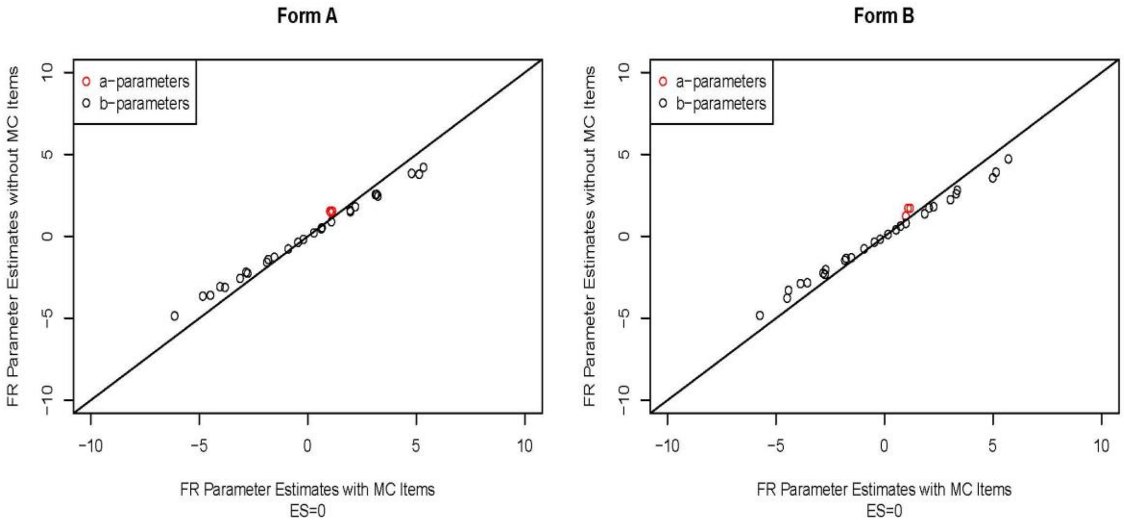


Figure 4.29 Comparison of English Language FR item parameter estimates for ES=0.

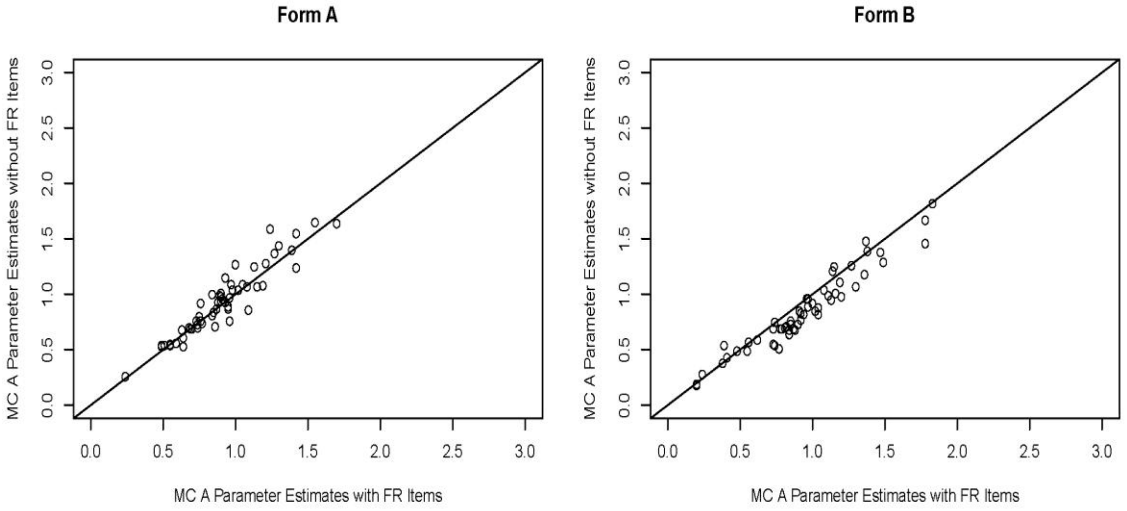


Figure 4.30 Comparison of French Language MC item *a*-parameter estimates for ES=0.

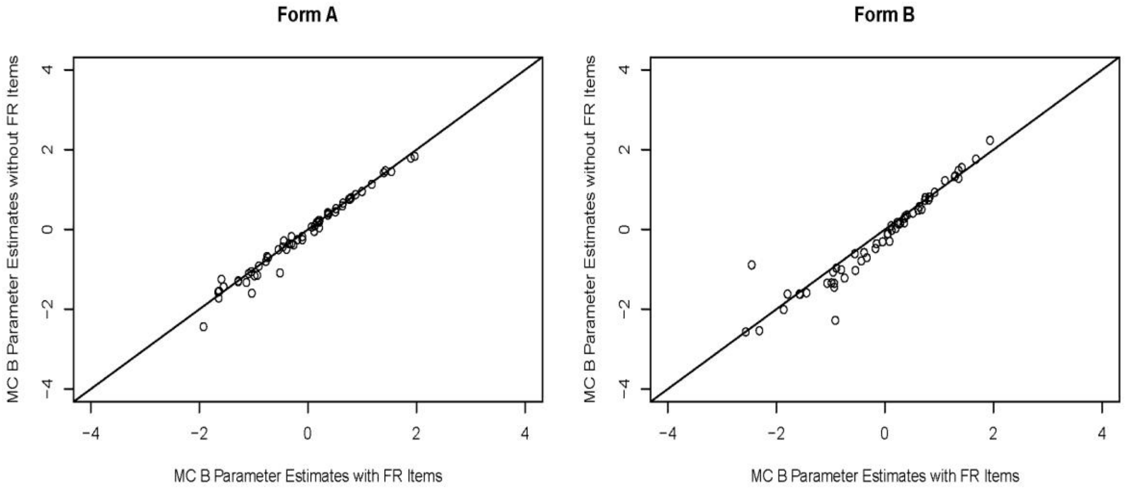


Figure 4.31 Comparison of French Language MC item *b*-parameter estimates for ES=0.

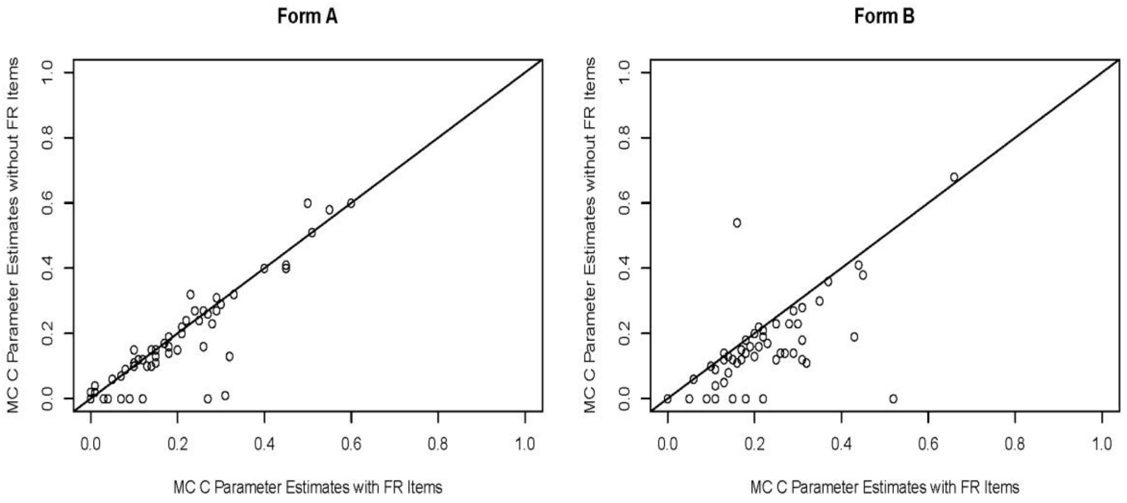


Figure 4.32 Comparison of French Language MC item *c*-parameter estimates for ES=0.

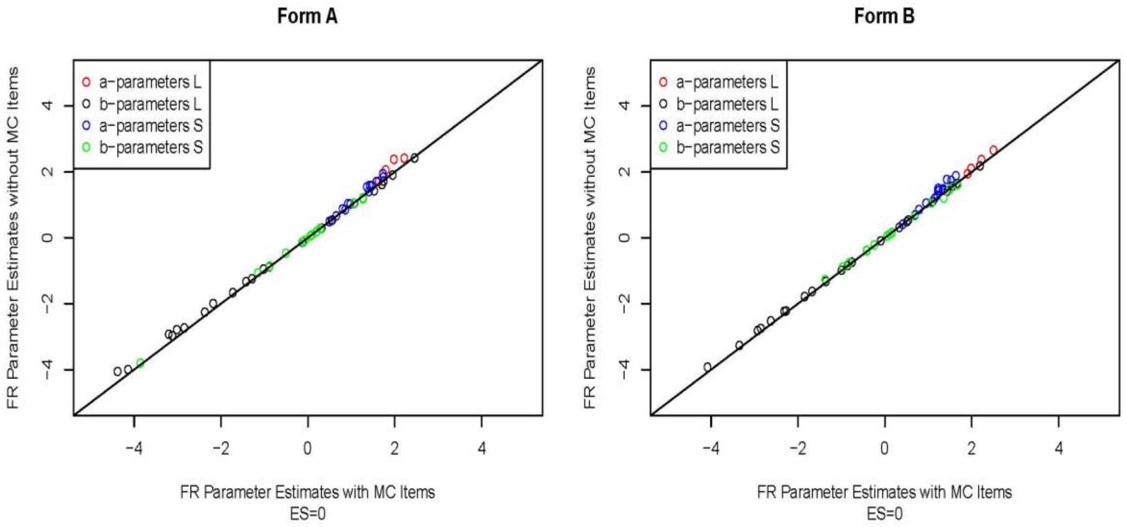


Figure 4.33 Comparison of French Language FR item parameter estimates for ES=0.

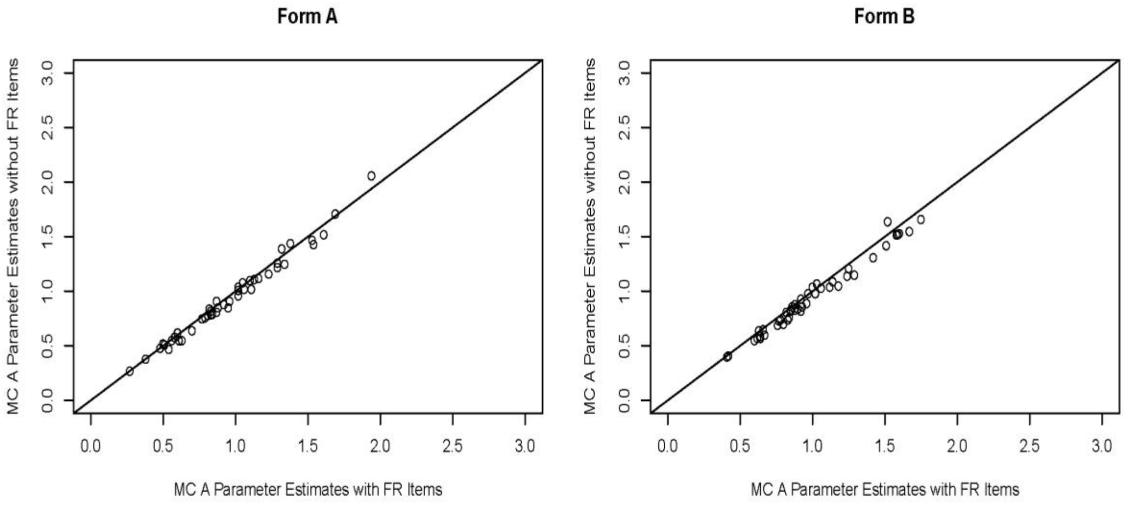


Figure 4.34 Comparison of Physics B MC item *a*-parameter estimates for ES=0.

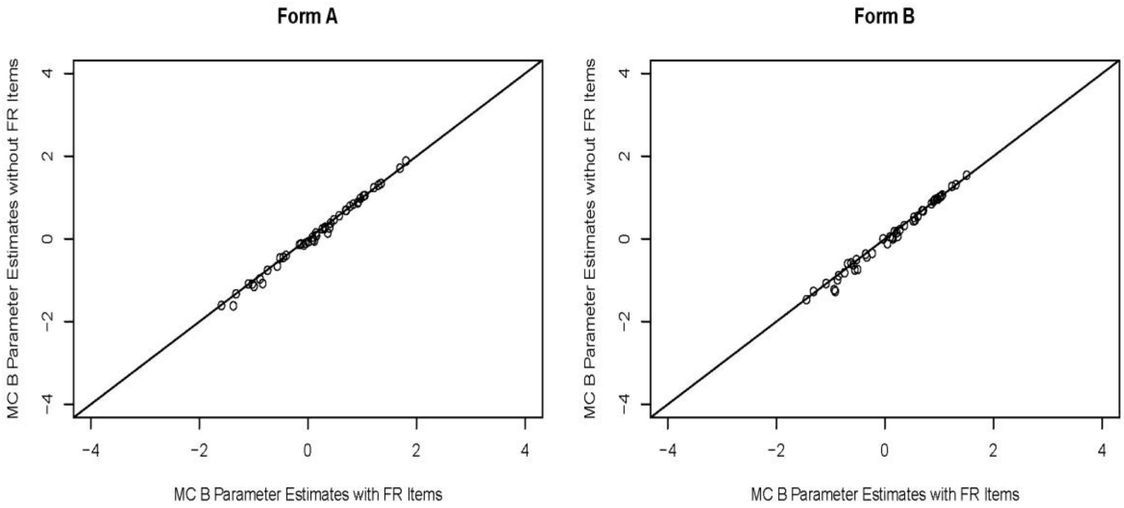


Figure 4.35 Comparison of Physics B MC item *b*-parameter estimates for ES=0.

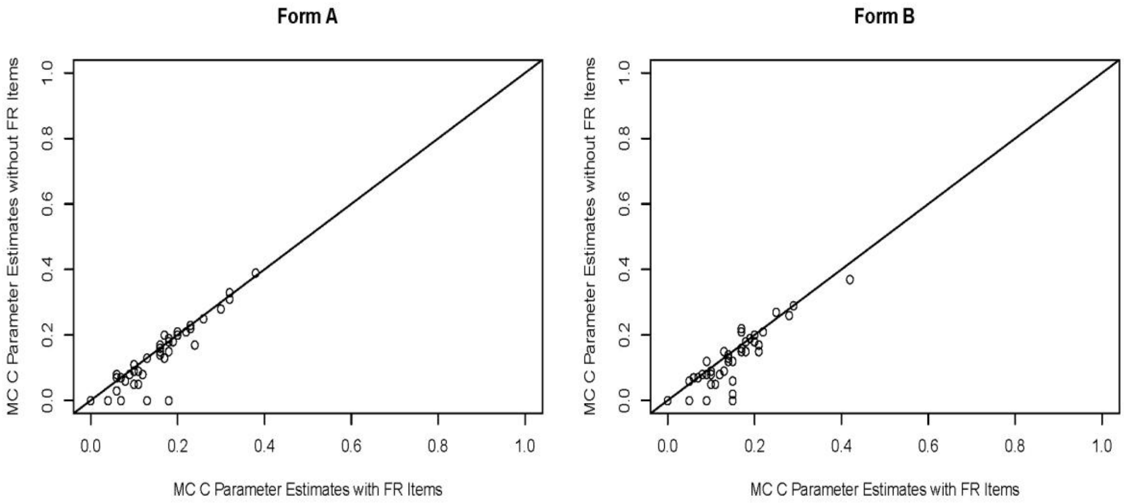


Figure 4.36 Comparison of Physics B MC item *c*-parameter estimates for ES=0.

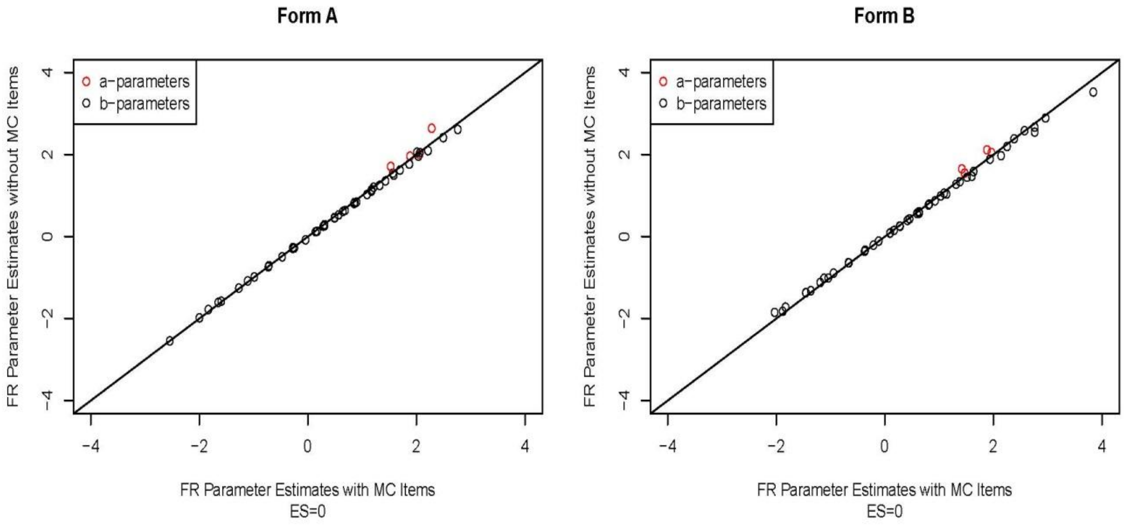


Figure 4.37 Comparison of Physics B FR item parameter estimates for ES=0.

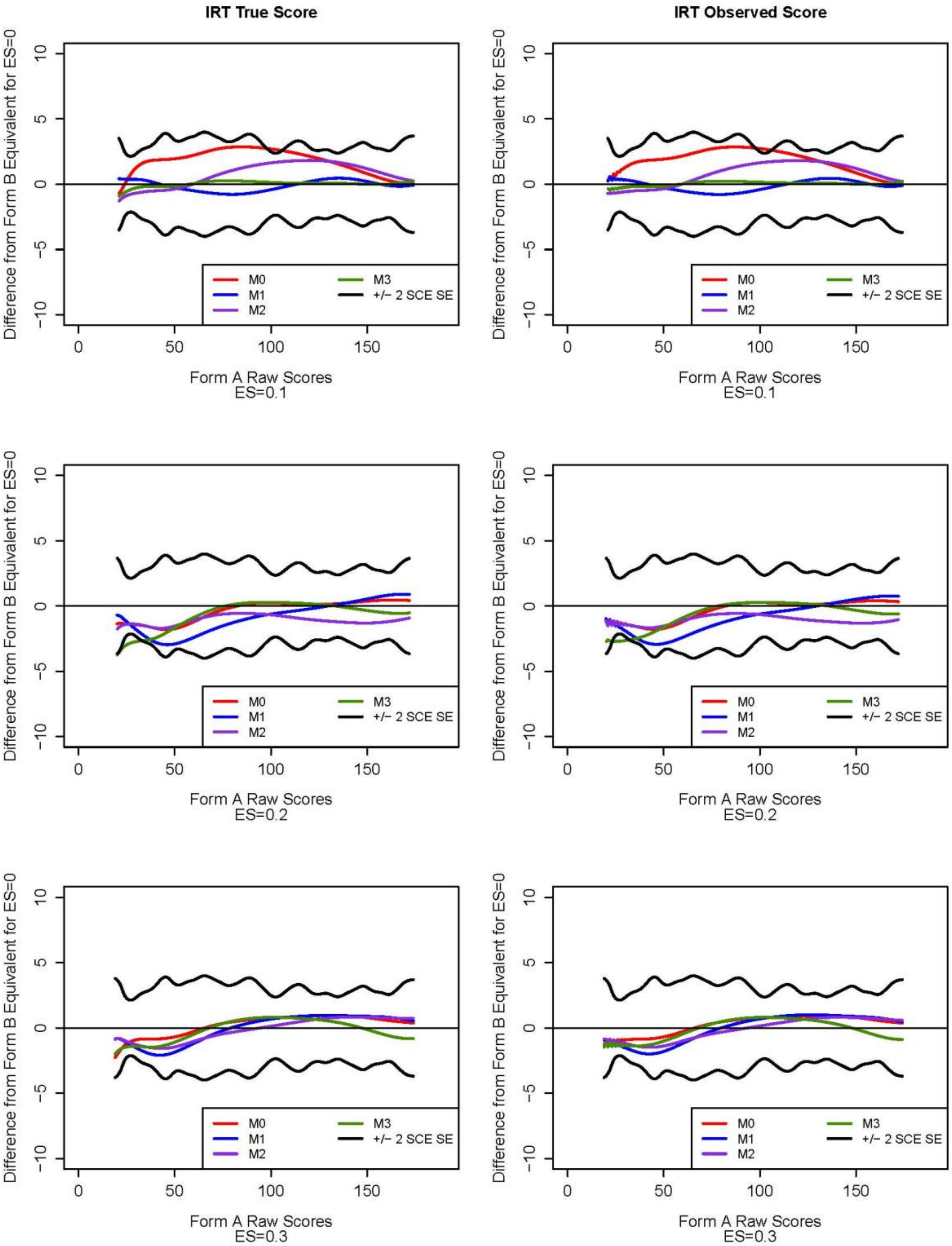


Figure 4.38 Comparison of Chemistry Matching Methods across ES Levels for IRT Equating Methods.

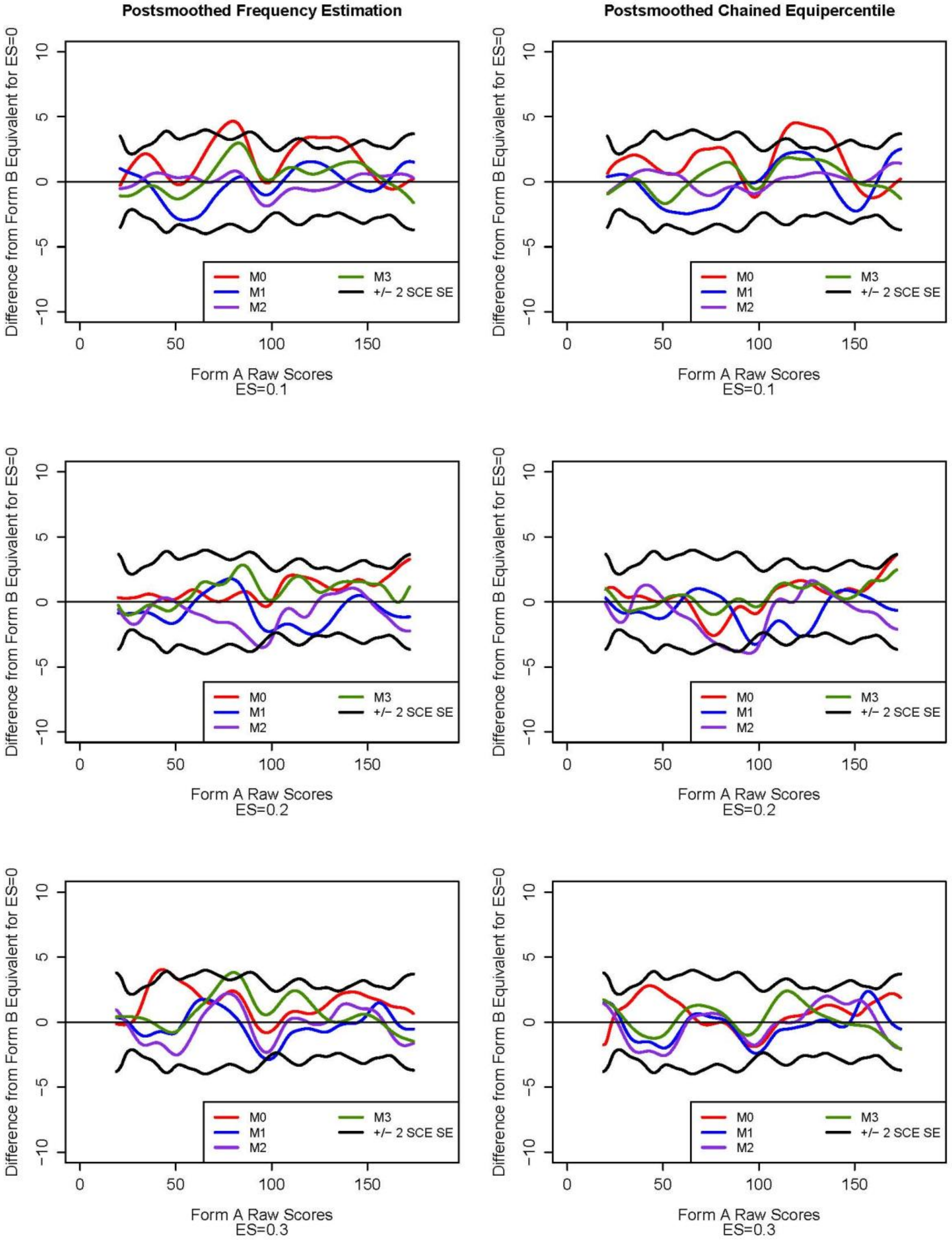


Figure 4.39 Comparison of Chemistry Matching Methods across ES Levels for Traditional Equating Methods.

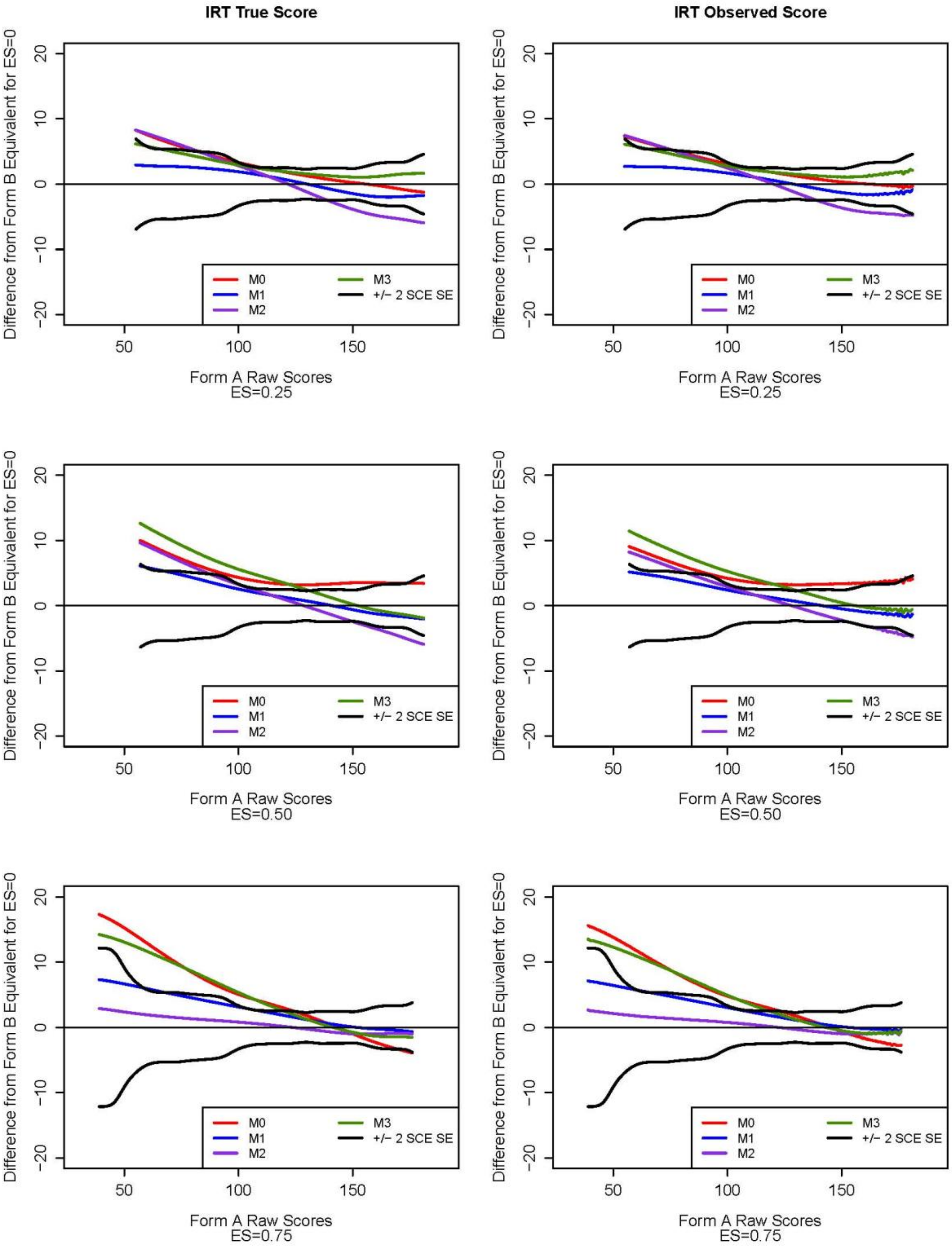


Figure 4.40 Comparison of English Language Matching Methods across ES Levels for IRT Equating Methods.

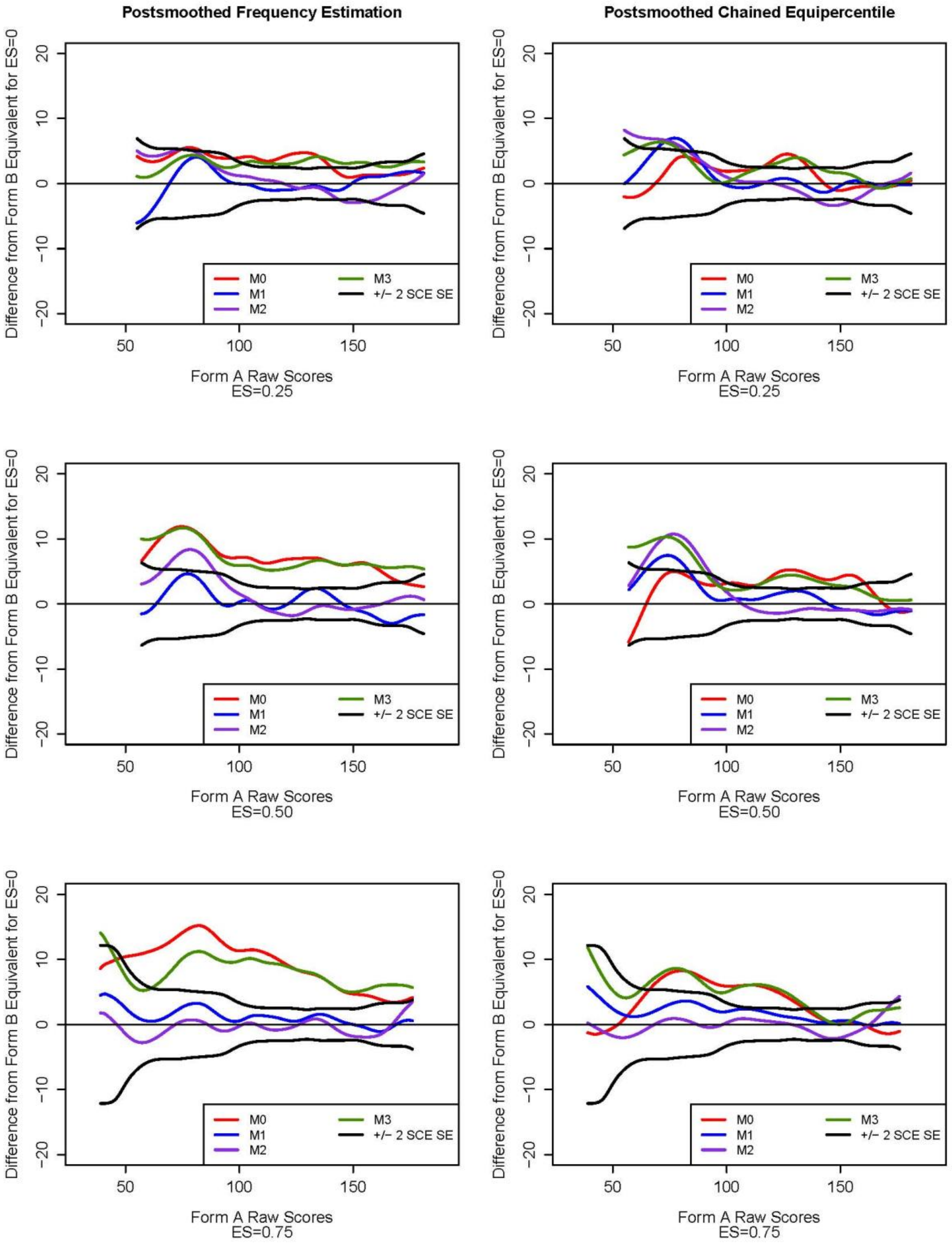


Figure 4.41 Comparison of English Language Matching Methods across ES Levels for Traditional Equating Methods.

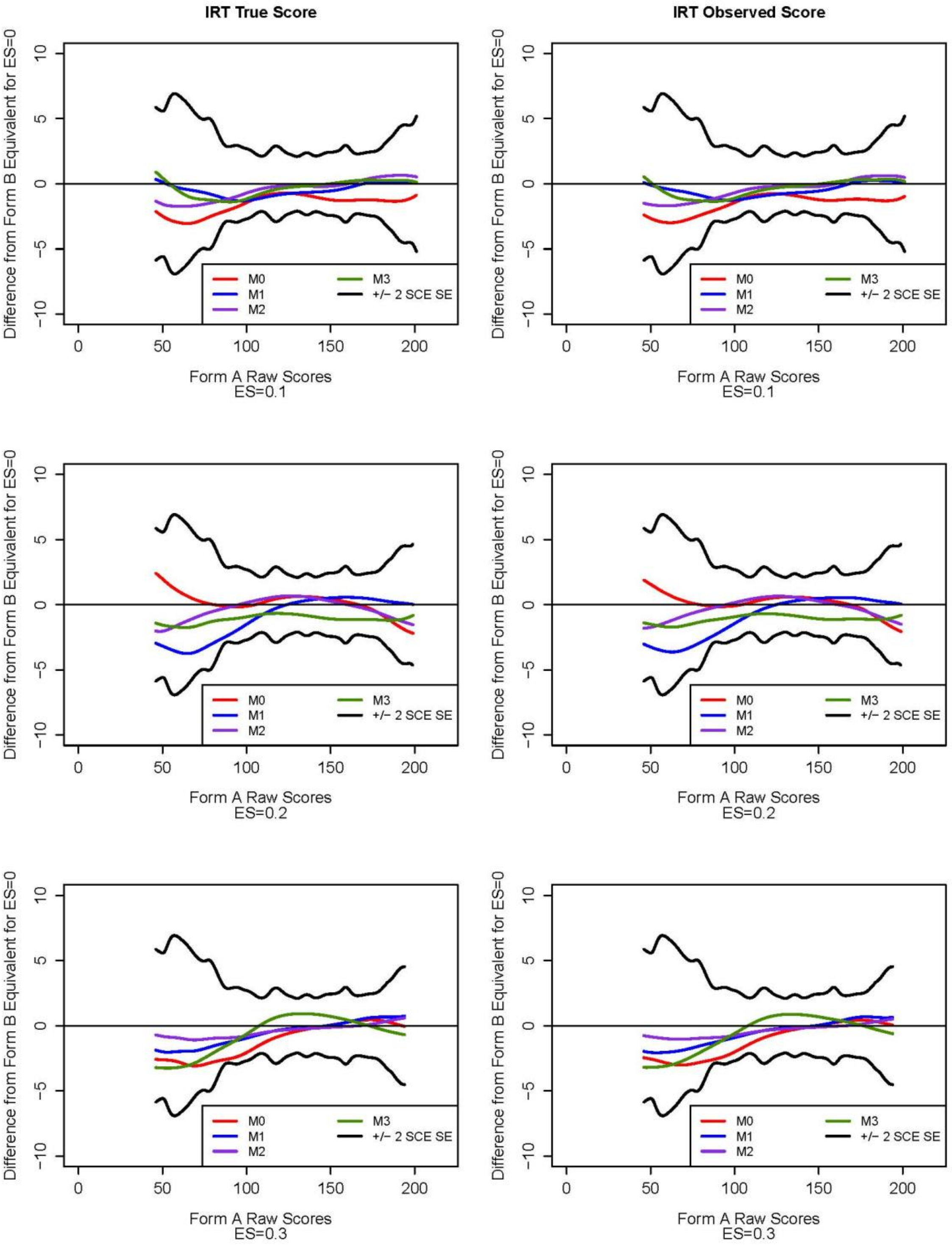


Figure 4.42 Comparison of French Language Matching Methods across ES Levels for IRT Equating Methods.

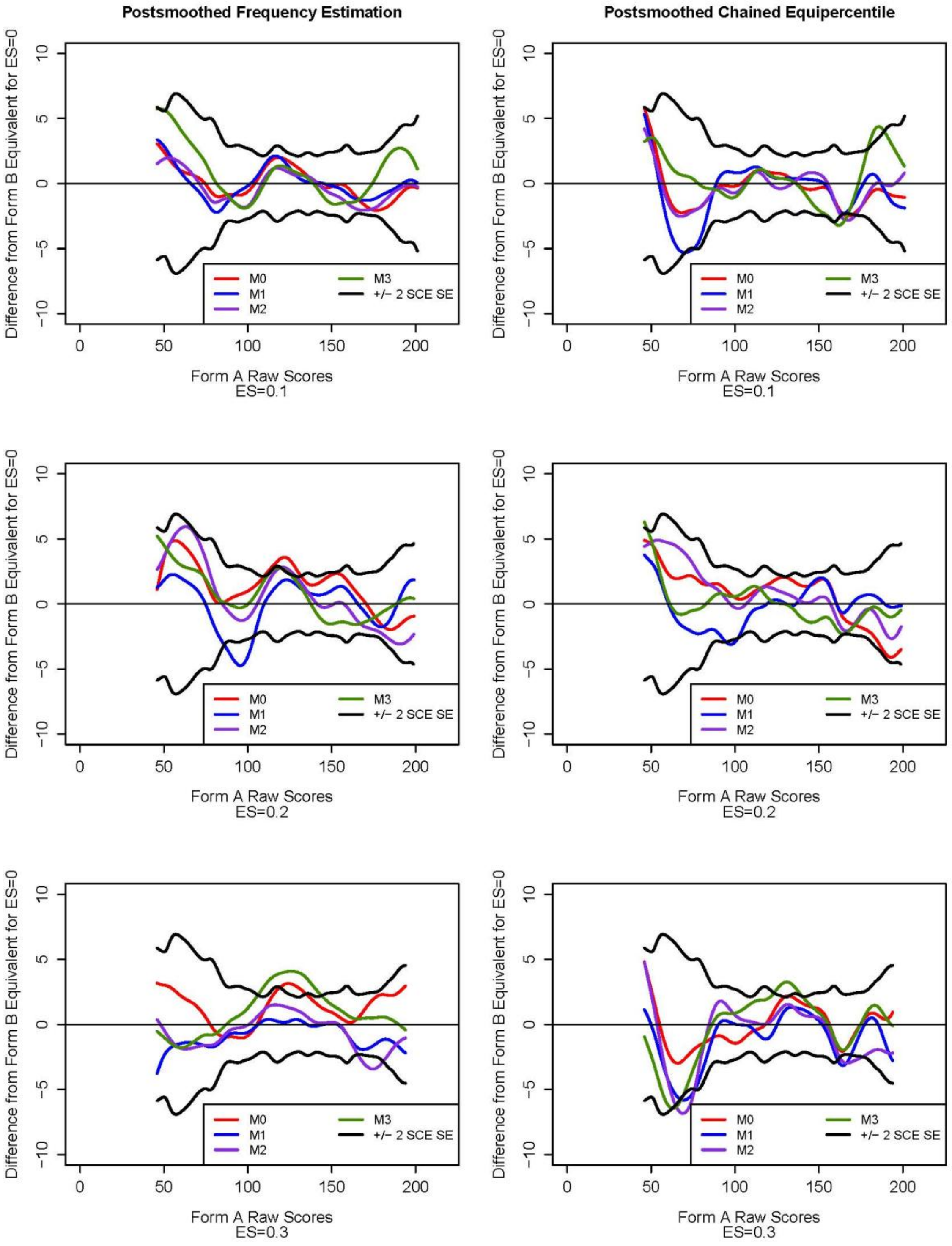


Figure 4.43 Comparison of French Language Matching Methods across ES Levels for Traditional Equating Methods.

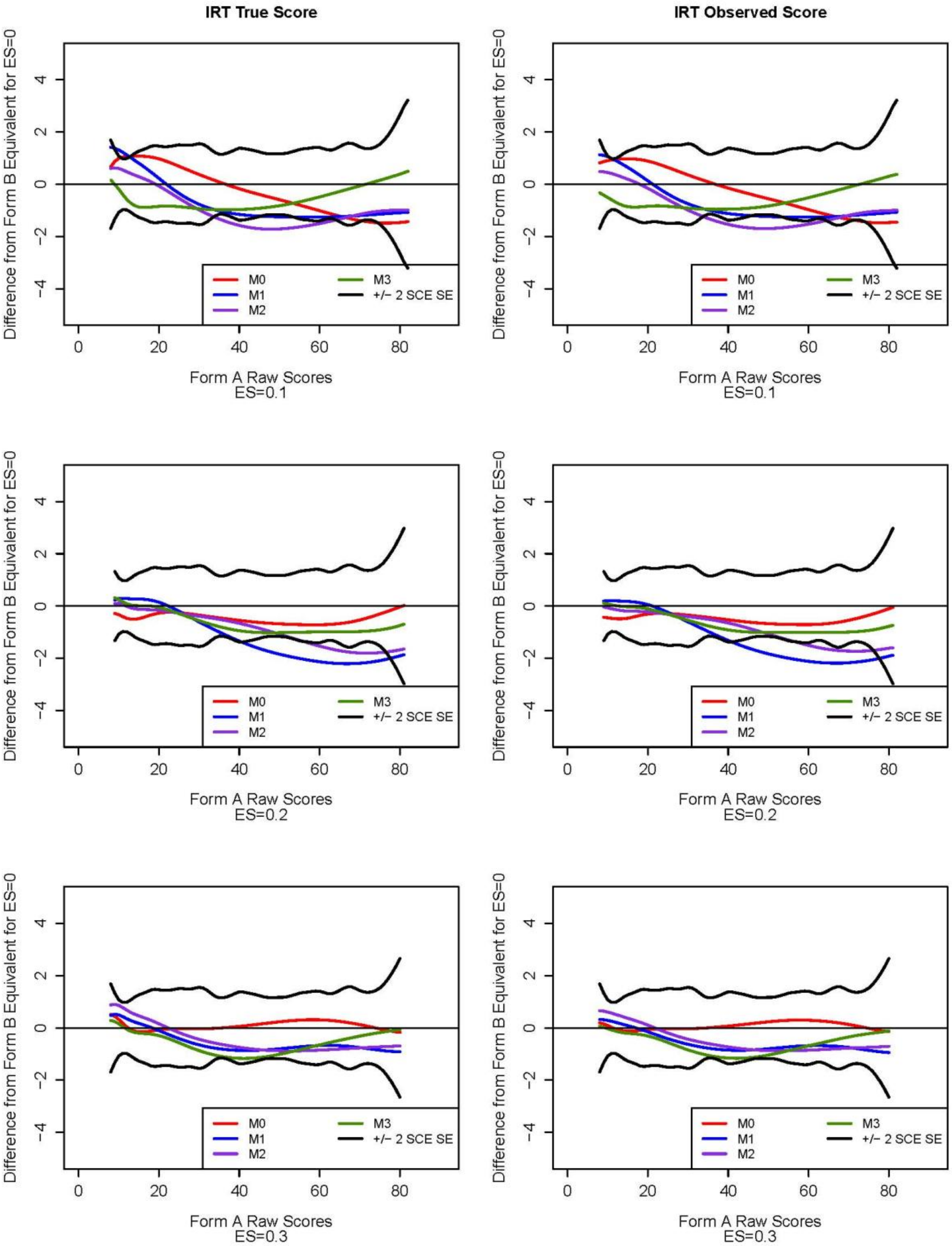


Figure 4.44 Comparison of Physics B Matching Methods across ES Levels for IRT Equating Methods.

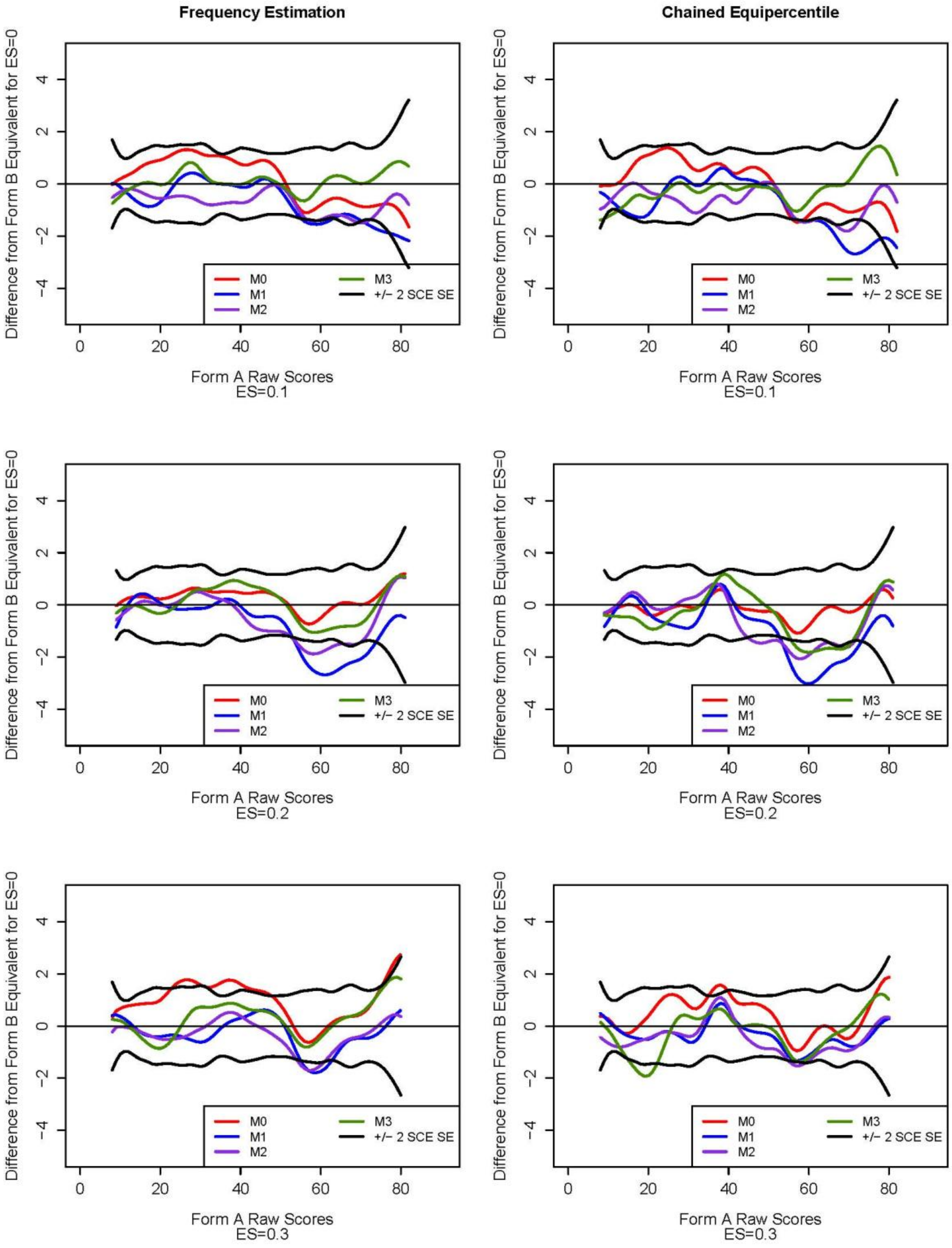


Figure 4.45 Comparison of Physics B Matching Methods across ES Levels for Traditional Equating Methods.

CHAPTER 5

DISCUSSION

This chapter includes a summary of the results described more thoroughly in Chapter 4. In addition, the significance of the study, limitations of the study, and future research is described. The last section provides conclusions based on the outcomes of this dissertation.

Summary of Results

The summary of results includes six sections, following the same presentation order used in Chapter 4. The first section provides a summary of the comparison of SG and CINEG equating results. The second section provides a summary of results from the population invariance analyses. The third section summarizes the equating results found with unmatched groups. The fourth section provides a summary of the findings related to the adequacy of equating assumptions with unmatched groups. The fifth section summarizes the results of matched samples equating, and the final section provides a summary of the adequacy of equating assumptions with matched samples. Where possible, the relationship of the findings to previous research is described.

SG versus CINEG Equating Results

An initial comparison of the SG and CINEG results using *Equating Recipes* indicated that the program provided reasonable results. The SG and CINEG unsmoothed traditional equating relationships were the same, as expected, given that the old and new form groups were the same set of examinees (i.e., the SG design). The smoothed traditional SG and CINEG equating results were very similar but differed slightly due to the way the bootstrap SEs were calculated for the two designs and used in cubic spline postsmoothing. The IRT SG and CINEG results differed only because the item parameters were estimated simultaneously for the SG methods and separately for the

CINEG methods. Equating methods used with the SG design have few statistical assumptions. The fact that SG and CINEG equating methods provide identical or nearly identical results for equivalent groups, indicates that CINEG equating results, used when the common item ES is zero, provide a baseline that should not be distorted by possible violations of equating assumptions.

Population Invariance

SG equating methods were used to assess population invariance of equating relationships for parental education level subgroups. The IRT equating relationships for each subgroup were within plus or minus two equipercentile SEs of the total group equating relationship for the majority of the score scale for all but Physics B, where the lowest parental education subgroup deviated by slightly more than two SEs for several score points. The unsmoothed equipercentile equating results for the subgroups often exceeded two SEs. Although the unsmoothed equipercentile SEs may not be accurate for the IRT equating methods, it appeared that the IRT equating methods were less population dependent than the traditional methods. IRT REMSD values were lower in general than equipercentile values, and often below the SDTM criteria. Classification consistency was higher for the IRT methods than for the equipercentile methods, indicating that the traditional equating results appeared more population dependent. However, classification consistency was higher than 95% across all equating methods and subgroups. In general, it appeared that the equating relationship for the lowest parental education subgroup deviated most from the total group equating relationship, but all equating relationships were within sampling error of the total group equating relationship for the majority of the score scale. Comparisons of equating relationships, REMSD values, and classification consistency values indicated that the subgroup equating relationships were fairly similar for all of the exams.

In previous research investigating the population invariance of AP Exam equating relationships (Dorans, 2003; von Davier & Wilson, 2006), population invariance did not appear to be a stable characteristic of an exam. The degree to which subgroup equating relationships differed from total group equating relationships depended on such characteristics as administration year, exam, equating method, and subpopulations considered. In this study, only one subpopulation and administration year was considered. However, across the four AP Exams considered, IRT true and observed score equating methods consistently showed less population dependence than the equipercentile equating method. In general, the equating relationships for parental education subgroups showed little population dependence. The use of the SG design in this dissertation allowed population invariance to be assessed without the contaminating effects of violations of statistical assumptions that are more likely to be found with the CINEG design. An important aspect of the population invariance analyses in this dissertation was the inclusion of SEs and classification consistency as evaluation criteria which have not been used in previous population invariance studies.

Equating with Unmatched Groups

Group performance was systematically altered so that old and new form groups differed by as much as 0.75 standard deviations. Traditional and IRT CINEG equating methods were used at different ES levels. A 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 high ES levels. Lord and Wingersky (1984) also found that the two IRT equating methods provide very similar results. The equating results for postsmoothed frequency estimation and chained equipercentile became less similar as ES increased, and the traditional and IRT results became less similar as ES increased. Similar findings were reported by Harris and Kolen (1990).

For the extreme ES values used with English Language, the deviation of comparison equating relationships (ES>0) from the criterion equating relationship (ES=0) increased as ES increased in terms of graphical comparisons, REMSD values, and classification consistency. However, for the three exams with smaller ES values, there did not appear to be a relationship between ES and the degree to which the comparison and criterion equating relationships differed. For ESs ranging from 0 to 0.3, most comparison equating relationships were within plus or minus two SEs of the criterion. IRT equating methods tended to have higher classification consistency values than the traditional equating methods across ES levels and exams. Physics B had the highest classification consistency values. IRT equating relationships were noticeably more smooth than the traditional equating relationships, especially for frequency estimation. The traditional results would have been smoother with a higher smoothing parameter. The criteria used to select an S-value in this dissertation may have been too conservative.

The comparison equating relationships at ESs of 0.5 and 0.75 were dramatically different from the criterion equating relationships for English Language, especially for the frequency estimation method. These differences could not be attributed to population invariance, because, as previously mentioned, the population invariance analyses indicated minimal population dependence for the four AP Exams. Therefore, the inaccuracy of the equating relationships based on groups that differed substantially in performance may have been caused by violations of equating assumptions. To test this hypothesis, assumptions were assessed for all traditional and IRT CINEG equating methods.

Evaluating Equating Assumptions for Unmatched Groups

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 to remain 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 equipercentile 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.   Although achievement tests like those used in this dissertation often reveal one dominant dimension, it seems likely that there are other small dimensions involved in the process by which examinees respond to test items. As group differences increase, these smaller dimensions may have a larger impact on IRT equating results.

Weighted absolute maximum difference statistics were used to assess the frequency estimation method assumptions. For all exams, the smaller the ES, the better the FE assumptions held. For English Language, there was some indication that the degree to which the equating assumptions held corresponded directly to the accuracy of the equating results as quantified using the REMSD statistic.

SG REMSD statistics were used to assess the chained equipercentile assumptions. Increases in ES corresponded to increases in REMSD for the extreme English Language ESs, but not for the more moderate ESs used with the other three exams. However, across all exams, the SG REMSD values corresponded directly to the chained equipercentile REMSD values, indicating that the degree to which chained equipercentile assumptions hold corresponds to the accuracy of chained equipercentile equating results.

Several methods were used to assess the unidimensionality assumption of the IRT equating methods. The MC-FR correlations indicated that the language exams had lower disattenuated correlations than the science exams. English Language had an especially low disattenuated correlation, indicating that the MC and FR items for this exam may not measure the same construct. Principal component analysis was also conducted and generally indicated one dominant dimension and one or two smaller dimensions for all four exams. The MC-FR correlations and principal component analysis results were consistent across ESs. An attempt was made to use PolyDIMTEST and PolyDETECT software to assess the dimensionality of the full-length operational exams, but PolyDIMTEST appeared overly sensitive to program specifications and PolyDETECT was not able to accommodate the FR sections of Chemistry and Physics B. There was some indication that the language tests were multidimensional and the science MC sections were unidimensional when assessed using PolyDETECT. Because of the limitations of these programs, they were not used with subforms A and B.

Finally, IRT item parameters were estimated for the MC items only, for the FR items only, and for MC and FR items simultaneously to assess the stability of item parameter estimates. MC item parameter estimates increased when estimated with FR items. FR *a*-parameter estimates decreased when estimated with MC items. As the ES increased, Form A *b*-parameter estimates for both MC and FR items increased whether they were estimated separately for each item type, or together; Form B *b*-parameter estimates for both MC and FR items decreased.

Based on the results of the assumption evaluations, it appears that large group differences between old and new form groups decrease the degree to which the assumptions hold for the traditional CINEG equating methods considered in this study. This trend held across all four exams. This finding helps to explain the general finding in the literature that group differences often lead to inaccurate and inconsistent equating results (Kolen, 1990).

Although a variety of model fit and dimensionality analyses were conducted, no discernable relationship was found between the magnitude of group differences and the degree to which IRT assumptions held. However, the accuracy of IRT equating decreased as ES 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 dissertation, especially since IRT single group equating results proved to be more population invariant than the traditional equating results.

Matched Samples Equating

Matching of old and new form groups was considered as a possible way to improve the accuracy and consistency of equating results. Four matching methods were considered: no matching (M_0), matching on the selection variable only (M_1), matching on a propensity score based on six variables including the selection variable (M_2), and matching on a propensity score based on five variables, not including the selection variable (M_3). The nonselection variables used in M_2 and M_3 were related to the selection variable to varying degrees. The fee reduction indicator and ethnicity had the highest phi-coefficient with parental education. US region and examinee grade level were slightly related to parental education, and gender was not related to parental education. Because of the relationship between some of the examinee background variables and parental education, pseudo r-squared values for M_3 were nonzero and increased as ES increased. Also, the coefficients for background variables other than parental education were significant for M_2 and M_3. However, the matched frequencies across levels of parental education were similar for M_1 and M_2, but the frequencies for M_3 were much closer to the unmatched (M_0) frequencies. Similarly, M_1 and M_2 resulted in ESs near zero even when the M_0 ESs were as high as 0.75. M_3 decreased the most extreme ESs for English Language somewhat, but in general did not result in matched ESs near zero. Therefore, the M_3 matching method did not provide a good approximation to matching on the selection variable. However, the variables included in the M_3 matching method were only moderately related to the selection variable. Propensity score matching with variables more closely related to the selection variable, like prior years' test scores (McClarty, Lin, & Kong, 2009; Way, Davis, & Fitzpatrick, 2006; Way, Lin, & Kong, 2008) might prove more effective..

Equating relationships for the three exams with smaller ESs were not noticeably improved by matching methods M_1-M_3. As with M_0, the results were mostly within plus or minus two SEs of the criterion equating relationship. For English Language, M_1 and M_2 provided closer results to the criterion at high ES levels compared to M_0 and M_3. The frequency estimation method appeared to be the most sensitive to group differences and the most improved by matching methods M_1 and M_2. For English Language the old form equivalent means were more similar across equating methods for M_1 and M_2 compared to the means for M_0 and M_3. Cook, Eignor, and Schmitt (1988, 1990), Lawrence and Dorans (1990), and Schmitt, Cook, Dorans, and Eignor (1990), also found that matching resulted in more consistent equating relationships across methods.

REMSD values tended to increase for M_0 as ES increased for the more extreme English Language ESs. However, M_1 and M_2 provided consistently smaller REMSD values only for the frequency estimation method. M_3 did not provide an improvement. IRT REMSD values were the lower than traditional REMSD values for M_0 and M_3. None of the matching methods lowered REMSD values below the SDTM criterion.

Classification consistency with the ES=0 equating relationship as the criterion was improved by M_1 and M_2 for English and French Language Exams, but not for Chemistry or Physics B. Matching methods M_1 and M_2 improved classification consistency dramatically for frequency estimation compared to the unmatched English Language ESs.

With IRT true score equating relationships as the criterion, classification consistency was very high for IRT observed score equating, but less high for the traditional methods. For M_0, classification consistency decreased as ES increased for the traditional methods. For English Language, M_1 and M_2 resulted in higher classification consistency for the traditional equating methods than for M_0 and M_3. There did not appear to be a relationship between classification consistency and matching method for the three exams with smaller ES values.

There was some indication with English Language that the frequency estimation method produced more biased equating results than the other equating methods when group differences were large. Sinharay and Holland (2007) and Holland, Sinharay, von Davier, and Han (2008) also found that the frequency estimation method is more biased than the chained equipercentile method. Ricker and von Davier (2007) found that the frequency estimation method can be less biased than the chained equipercentile method when the common item set is small relative to the total test length. However, the common item sets in this dissertation were proportionally larger than they often are in practice because of the way Forms A and B were constructed. Therefore, the finding that the chained results were less biased in this dissertation corroborates previous research findings. The frequency estimation method appeared less sensitive to sampling technique than the chained or IRT methods in some of the matched samples equating studies conducted by ETS, but not in others (Cook, Eignor, & Schmitt, 1988, 1990; Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990; Wright & Dorans, 1993). A possible explanation for this inconsistency is that the ES levels considered in most of the ETS research were no larger than 0.4. In this dissertation it was found that the relationship between equating accuracy and ES was only consistent when the ESs were very large. For smaller ES values, like those used in the ETS matched samples studies, the relationship between ES and equating accuracy appears to be obscured by the magnitude of random equating errors. When group differences are small, the frequency estimation method may actually have less total equating error compared to the chained equipercentile method because the frequency estimation SEs are smaller than the chained SEs. Smaller frequency estimation SEs were also noted by Sinharay and Holland (2007) and Wang, Lee, Brennan, and Kolen (2008).

The choice between traditional methods should take into consideration both the magnitude of random equating error and the size of group differences. In many operational testing situations, group differences may be small and the frequency estimation method would be the preferable equating method from the perspective of total equating error. However, when large group differences are anticipated, if matched sampling is not used, the chained equipercentile method may provide optimal results. With large group differences, matched sampling greatly reduces the equating bias of the frequency estimation method, making it again a reasonable choice. Choice between the traditional and IRT methods is more complex, involving random and systematic equating errors, the dimensionality of the assessments, and other considerations beyond equating design.

Evaluating Equating Assumptions with Matched Samples

The degree to which equating assumptions held was assessed for the frequency estimation method using the weighted absolute maximum difference between cumulative distribution functions for the conditional distribution of Form A composite scores given common item scores, and for the conditional distribution of Form B composite scores. The weighted absolute maximum differences increased as ES increased for M_0 but did not increase for M_1 and M_2. The weighted absolute maximum differences also tended to be smaller for M_1 and M_2. M_3 only provided an improvement over M_0 for English Language. Therefore, matching on the selection variable resulted in better fit of the data to frequency estimation assumptions. Also, a comparison of weighted absolute maximum differences to frequency estimation REMSD values indicated that the degree to which frequency estimation assumptions held corresponded to the accuracy of equating results for the extreme ESs of English Language.

For the chained equipercentile equating method, assumptions were evaluated using SG REMSD statistics for the relationship between composite scores on Form A and common items scores, and for the relationship between the common item scores and the composite scores on Form B. As previously mentioned, REMSD values increased as ES increased for English Language, but not for the other exams. M_1 and M_2 provided smaller REMSD values than M_0 for Chemistry and English Language in most cases, but did not consistently decrease REMSD for French Language or Physics B. When SG REMSD statistics were compared to CINEG REMSD statistics for the chained method, increases in SG values tended to correspond to increases in CINEG values (although the pattern was not consistent for Physics B). The relationship of SG and CINEG REMSD values indicates that the degree to which chained equipercentile equating assumptions hold corresponds to the accuracy of chained equating results.

The unidimensionality assumption for the IRT equating methods was assessed using correlational analyses and comparisons of item parameter estimates. A description of the correlational analyses was provided previously for unmatched groups. Matching methods did not change the results. However, for item parameter estimates, there were some differences for matched (M_1-M_3) and unmatched (M_0) groups. MC *a*-, *b*-, and *c*-parameters tended to increase when estimated with FR items for matched and unmatched groups. Also, the FR discrimination values decreased when estimated with MC items for both unmatched and matched groups. However, with unmatched groups, the MC and FR *b*-parameter estimates tended to increase as ES increased for Form A, and the *b*-parameter estimates tended to decrease as ES increased for Form B. This pattern was found whether the MC and FR items were estimated separately or simultaneously. For matched groups, this pattern was less consistent, and in some cases was not found at all. The increase in *b*-parameter estimates for lower performing groups is an artifact of the scaling used in MULTILOG that centers ability at zero with a standard deviation of one. With matching, the groups are more similar in performance and the *b*-parameters do not systematically increase for the old form group or decrease for the new form group.

Summary of Study Findings and Related Research

In this dissertation it was found that:

- IRT equating methods appeared less population dependent than traditional methods.
  - However, Dorans (2003) and von Davier and Wilson (2006) compared traditional and IRT equating results and found that the results indicated a similar degree of population invariance.
  - This dissertation included SEs and classification consistency as population invariance evaluation criteria. These criteria have not been used in previous population invariance studies.
- IRT true and observed score equating methods produced very similar results even at high ES levels.
  - Lord and Wingersky (1984) also reported that IRT true and observed score equating results were very similar.
- The frequency estimation method had less random equating error than the chained equipercentile method.
  - These results have been previously reported by Sinharay and Holland (2007) and Wang, Lee, Brennan, and Kolen (2008).
- Frequency estimation and chained equipercentile results became less similar as ES increased and the traditional equating results became less similar to IRT equating results for extreme ESs.
  - Harris and Kolen (1990) also found that frequency estimation and chained equipercentile results diverge as ES increased.
  - Equating results were also shown to diverge for traditional and IRT equating methods when old and new form groups differed substantially by researchers at ETS (Cook, Eignor, & Schmitt, 1988,1990; Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans,

    & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990; Wright & Dorans, 1993).

- As group differences increased equating accuracy decreased, but equating results did not consistently exceed sampling error for ESs less than 0.3.
    - The link between large group differences and equating accuracy has not been studied systematically in previous studies. These findings are new and require replication.
- The frequency estimation method appeared most sensitive to large group differences.
    - This sensitivity of the frequency estimation method to group differences has been reported by Holland, Sinharay, von Davier, and Han (2008), Ricker and von Davier (2007), and Sinharay and Holland (2007).
- For traditional equating methods, the higher the ES, the less well the equating assumptions held.
    - For the traditional equating methods the degree to which the equating assumptions held corresponded to the accuracy of the equating results.
    - Based on the methods used to assess dimensionality in this dissertation, a relationship between the ES and the equating assumptions was not found for IRT equating methods.
    - The relationship between group differences, equating assumptions, and equating accuracy has not been studied systematically in other studies. These findings are new and require replication.
- Matching methods using the selection variable were successful in decreasing group differences, but matching on a set of variables only modestly related to the selection variable was not successful.
    - Matching on the selection variable decreased group differences in a study by Livingston, Dorans, and Wright (1990).

  - Matching did not substantially improve group differences in a study by Yu, Livingston, Larkin, and Bonett (2004) where the selection variable was not known.
- For extreme ESs, matching methods using the selection variable resulted in more accurate equating results for all equating methods (for ESs less than 0.3, the results were within sampling error of unmatched results).
  - Matched samples studies by ETS used common item scores instead of the selection variable and did not find good results (Cook, Eignor, & Schmitt, 1988,1990; Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990; Wright & Dorans, 1993)
  - These findings are new and require replication.
- Frequency estimation results were most improved by use of matched sampling.
  - Previous findings about the sensitivity of the frequency estimation method to sampling technique were mixed, possibly due to the use of common items instead of the selection variable(s) and/or the modest ESs considered in the studies (Cook, Eignor, & Schmitt, 1988,1990; Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990; Wright & Dorans, 1993).
- For traditional equating methods, matching on the selection variable did improve the degree to which equating assumptions held for extreme ESs.
  - Based on the methods used to assess dimensionality in this dissertation, a relationship between matching methods and the equating assumptions was not found for IRT equating methods.
  - Previous studies have not compared the impact of matched sampling on equating assumptions. These results are new and require replication.

- Matching methods using the selection variable resulted in more consistent equating results across methods for extreme ESs.
    - Similar findings were reported by Cook, Eignor, and Schmitt (1988, 1990), Lawrence and Dorans (1990), and Schmitt, Cook, Dorans and Eignor (1990).

Significance of the Study

Research has shown that group differences can cause inaccurate and inconsistent equating results (Kolen, 1990). Matched sampling has been employed in several studies as a potential solution to the equating problems encountered with group differences (e.g., Lawrence & Dorans, 1990; Way, Davis, & Fitzpatrick, 2006; Yu, Livingston, Larkin, & Bonett, 2004). However, none of the existing research has investigated the cause of equating inaccuracies when groups differed. This dissertation 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 appropriate matching techniques or other 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 strict 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 research has focused primarily on the DTM criterion for assessing population invariance (e.g., Dorans, 2003; von Davier & Wilson, 2006). In this study SEs were used as a way to distinguish large 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 exams and subgroups considered in this dissertation, it was clear that violations of equating assumptions were the main cause of inaccurate 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.

A variety of ESs were considered in this study, ranging from 0 to 0.3 for three of the exams, and from 0 to 0.75 for English Language. All of the main operational AP Exams had group differences of less than approximately 0.15 standard deviations from year to year. However, group differences might be larger for alternate AP forms that are used for some exams in cases where examinees have conflicts on the day the main form is given. Even with multiple yearly administrations, and administration groups that are known to differ in performance, the highest SAT ES in the Lawrence and Dorans (1990) study was only approximately 0.40. More extreme group differences including ESs larger than one have been found in some comparability studies (Yu, Livingston, Larkin, & Bonett, 2004). However, the ES ranges used with Chemistry, French Language, and Physics B may represent a range of ESs that would be found in typical equating situations. English Language provides a more extreme range of group differences so that the impact of group differences on equating results can be seen more clearly. The finding that even unmatched equating relationships are nearly within sampling error of the criterion equating relationship for ESs ranging from 0 to 0.3 may provide some reassurance that equating results may be reasonable even with moderate group differences. Only with ESs as large as 0.5 and 0.75 were the deviations of equating results from the criterion well beyond sampling error.

Matched sampling has been considered in previous research as a possible way to improve the accuracy of equating results when groups differ substantially (e.g., Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Paek, Liu, & Oh, 2006). One purpose of this dissertation was to investigate how matched samples equating could improve equating results. The methods used in this dissertation allowed the effects of population dependence to be disentangled from the effects of equating assumption violations. Because violations of equating assumptions appeared to have a greater impact on equating accuracy than population dependence, the impact of matched sampling on violations of equating assumptions was investigated. Based on comparisons of the statistical assumptions at various ESs, it appears that matching on the selection variable when group differences are very large improves the degree to which traditional equating assumptions are met, thereby increasing the accuracy of equating results. For IRT equating methods, although a direct link between group differences and equating assumptions could not be established, matching on the selection variable still improved equating accuracy in terms of producing equating results that were closer to the criterion equating relationship.

Additionally, while other studies have used matched samples equating (e.g., Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Paek, Liu, & Oh, 2006), and even used matched sampling operationally (Way, Davis, & Fitzpatrick, 2006; Way, Lin, & Kong, 2008), the methods used in this dissertation made it possible to match using the true selection variable. Because the selection variable was known, it was possible investigate how different matching methods would affect improvements in the degree to which equating assumptions hold and the accuracy of equating results. Although matching on only the selection variable (M_1) was considered the best-case-scenario for matched samples equating, an important finding was that matching on a propensity score where the selection variable was included in the logistic regression (M_2) provided comparable results. Using a propensity score including variables that are not selection variables is more likely to be the best-case-scenario in practice where the selection variable or set of variables is typically unknown.

The finding that matching on a propensity score that did not include the selection variable (M_3) did not provide improvement in the accuracy of equating results is also important. M_3 included variables that ranged from low to moderate in their relationships with the selection variable, and yet the matching was not successful in eliminating group differences, improving the degree to which equating assumptions were met, or increasing the accuracy of equating results. However, because the relationship between M_3 and parental education was so modest, M_3 was almost a worst-case-scenario for matching with variables that do not include the selection variable. Researchers at PEM have used propensity score matching with prior years' test scores (McClarty, Lin, & Kong, 2009; Way, Davis, & Fitzpatrick, 2006; Way, Lin, & Kong, 2008). Although the analyses used in this dissertation cannot address the efficacy of using prior test scores as matching variables, this seems like a promising method that may correlate much higher with theoretical selection variables.

Shadish, Cook, and Campbell (2002) warn that inappropriate matching can increase bias. However, there was little evidence that using M_3 had detrimental effects on equating results. Therefore, it appears that propensity score matching can be beneficial, especially with very extreme group differences, when the selection variable or set of variables is included in the logistic regression. However, if the propensity score does not happen to include the unknown selection variable(s), the matching process may not produce appreciably more or less accurate results than unmatched equating results. Variables that are very highly correlated with the selection variable, for example, previous years' test scores on related measures, may provide improved results, but additional research is necessary to confirm this hypothesis.

Finally, four different curvilinear equating methods were considered in this study: frequency estimation, chained equipercentile, IRT true score, and IRT observed score equating. Consistent with other studies (Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990), the consistency of equating results decreased as ES increased. Also, it appears that IRT true and observed score equating results are nearly identical even when the ES is very large. Lord and Wingersky (1984) also reported that the IRT true and observed score equating methods provided very similar results. This finding may be explained by the fact that equating accuracy appears to be related to the degree to which equating assumptions are violated, and both IRT methods share the same unidimensionality assumption. The IRT and chained equating results appear to be less sensitive to group differences than the frequency estimation method. Additionally, the accuracy of the frequency estimation method appears to be improved by matching (M_1 and M_2) more than the IRT and chained equating methods. As mentioned previously, findings about the sensitivity of equating methods to sampling method have been mixed (Cook, Eignor, & Schmitt, 1988, 1990; Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmittt, Cook, Dorans, & Eignor, 1990; Wright & Dorans, 1993). However, previous studies involved fairly modest group differences, which may have obscured the relationship between ES and equating accuracy.

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. For each exam, only one sample was selected for each ES. Sampling error was large enough that it was difficult to draw conclusions about the relationship between group differences and equating accuracy when the ES was less than 0.3. Inclusion of additional samples may have helped to disentangle the effects of sampling error from true differences in equating relationships.

In addition, a number of modifications were made to the scores and groups that may limit the generalizability of the results. Though the purpose of this dissertation was to investigate the relationship of group differences and equating accuracy in general, all of the data came from four AP Exams. Moreover, operationally, examinee responses are formula scored for AP Exams, but an imputation procedure was used to simulate number correct scoring. In addition, full-length operational forms were divided into two subforms (A and B) in order to have a SG equating design and criterion. Also, MC and FR items were weighted using integers instead of the operational noninteger section weights. After these substantial modifications, Form A and B scores are based on real data, but they do not resemble the original AP Exams in terms of test-length, composite to common item ratios, or scoring method.

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 selection 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.

In fact, the concept of the selection variable may make more sense in quasi-experimental design where person characteristics are likely to influence one’s treatment selection. For example, there are a variety of gender differences that would lead males to select some activities in greater proportions than females. However, the concept of the selection variable is much less straightforward in educational measurement. Often the only obvious reason that examinees choose different administration dates is because of their age or grade level. For example, because most examinees take the SAT in their 11th or 12th year of high school, the group of examinees that takes the SAT in a given year differs from groups of examinees that have taken the SAT in other years only in terms of their date of birth. However, the SAT is administered at multiple times of the year and administration groups are known to differ in performance. Possible selection variables might include the high school an examinee attended, the region of the US the examinee resides in, the "ambitiousness" of the examinee, whether or not the examinee plans to go to college, etc. In the case of the SAT there is likely research that would provide clues about the most likely selection variables. For other large-scale assessments, it is less clear why groups differ from administration to administration.

In practice it is likely that there are several variables that explain group differences, rather than the single selection variable used in this study. Moreover, many of the "true" selection variables may be difficult or impossible to measure. Even if the selection variables could be measured, there is always limited administration time with which to measure the variables.

Parental education was chosen as the selection variable 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. It appears that the magnitude of group differences is what determines the degree to which equating assumptions hold, and therefore the accuracy of equating results. Equating using a female old form group and a male new form group, for example, may not cause inaccurate equating results unless the common item ES between the two groups is large. However, this hypothesis cannot be substantiated given the results of this dissertation. If the common item ES is the real cause of assumption violations and inaccurate equating results, it seems reasonable to match groups based on common item scores only. A limitation of this dissertation is that a common item matching method was not considered. However, researchers at ETS have used common item matching with little success (see *Applied Measurement in Education, Volume 3*, (Special) Issue 1). The clear benefit of a common item matching method is that common item scores are available for all examinees. No additional administration time is needed. In addition, the matched samples can be equated using SG design equating methods.

This dissertation 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 dissertation. Representative common item sets are known to produce less biased equating results (Kolen, 1990). The frequency estimation and chained equipercentile assumptions involve the common items. A change in their composition would likely change both the adequacy of the equating assumptions, and the accuracy of equating results. However, the inclusion of FR items in the common item set can also be problematic because of security concerns and rater drift (Kim, Walker, & McHale, 2008).

Issues of composite-to-common item ratios were also not considered. Also, linear equating methods were not investigated in this dissertation. Additionally, only unidimensional IRT equating models were used despite having two item types and evidence that the language exams may be multidimensional. Also, all of the exams considered in this study involved MC testlets but special modeling was not done to ensure local independence. Finally, only one smoothing value was considered in this dissertation 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 dissertation may have been too stringent. Conclusions may be impacted by the degree of smoothing chosen.

Future Research

Because of the limited scope of this dissertation, there are several areas for future research. As mentioned in the last part of the limitations section, several more equating models could be considered including traditional linear equating methods, multiple smoothing values, and multidimensional IRT methods.

Including additional exams with large ESs would improve the generalizability of the findings presented in this dissertation, 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 ES and equating accuracy is stable across exams.

Inclusion of more samples for each ES level and matching method 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 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. The composition of the common item set is likely to affect both the adequacy of equating assumptions, and the efficacy of matched samples equating in improving equating accuracy.

The concept of the selection variable should also be given much more attention. More realistic selection variable(s) should be identified and used in future research. Consideration of possible matching variables should include the likelihood that groups differ on the particular variable, the relationship between the variable and exam scores, the feasibility of measuring the variable, and the improvement in accuracy achieved by matching with the variable. For example, common item score matching appears reasonable in that it captures administration group performance differences. The common item score is easy to match on because no additional measurement is required beyond the items already administered in the exam. However, research at ETS has called into question the efficacy of matching on common item scores in terms of reproducing criterion equating relationships (e.g., Schmitt, Cook, Dorans, & Eignor, 1990). The results of this dissertation suggest that more research on common item score matching is merited because the ESs reported in the ETS research were relatively small. The usefulness of common item score matching may increase when ESs are larger. Also, much of matched samples research at PEM has included previous years' test scores. When these types of scores are available, they appear convenient to use and may capture the selection variables of interest. The relationship of selection variables to group performance differences should be investigated, and if possible related to violations of equating assumptions and population invariance. Population dependent equating relationships may prove to be more of a problem in other contexts.

Finally, additional research about conditions that are likely to decrease the degree to which traditional and IRT equating assumptions hold should be identified. Methods to improve the degree to which equating assumptions are met might reveal optimal matching and equating methods.

Conclusions

As old and new form groups become less similar, CINEG equating assumption violations increase. When the degree to which equating assumptions hold decreases, the accuracy of equating results decreases. Subgroup equating relationships appear to be quite similar (i.e., population invariant) even when the subgroup performance differs substantially. This may be due to the fact that each subgroup equating relationship involves identical or nearly identical groups of examinees. These findings suggest that group dependence found when using CINEG equating methods (e.g., Dorans, 2003; von Davier & Wilson, 2006), may be caused by violations of equating assumptions, and not because the equating relationships differ from group to group.

The accuracy of equating results can be improved by matched samples equating when the matching technique incorporates the selection variable. If the selection variable is not known, and matching fails to include the selection variable(s), the accuracy of equating results may not be improved by matching. Additionally, the relationship between group differences, equating accuracy, and matching efficacy increases as ES increases. For ESs less than 0.3, the usefulness of matching may be questionable because the equating relationships may be within sampling error of the true relationship.

The frequency estimation method appears most affected by group differences, and also appears to benefit most from matching on the selection variable in terms of equating accuracy. The IRT and chained equipercentile equating methods appear to be less sensitive to group differences compared to the frequency estimation method. The consistency of equating results across methods decreases as group differences increase. However, IRT true score and observed score methods provide nearly identical results regardless of the magnitude of group differences.

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

REFERENCES

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes*. [Software version 1.0]. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment.

Census Regions and Divisions of the United States. Retrieved July 5, 2009 from www.census.gov/geo/www/us_regdiv.pdf

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*(1), 31-45.

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

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

D’Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*(1), 3-17.

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. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43-68.

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*(4) 281-306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). *Invariance of score linking across gender groups for three Advanced Placement program examinations* (ETS Research Report RR-03-27). Princeton, NJ: Educational Testing Service.

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*(1), 37-52.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 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*(2), 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*(1), 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*(1), 17-43.

Kim, H. K. (2006). *The effect of repeaters on equating: A population invariance approach.* Unpublished doctoral dissertation: University of Iowa.

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), 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*(1), 97-104.

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

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (ETS Research Report RR-88-23). Princeton, NJ: Educational Testing Service.

Lawrence, I. M., & Dorans, N. J. (1990). Effects on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*(1), 19-36.

Li, H.-H., & Stout, W. F. (1995). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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*(1), 73-95.

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

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

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.

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

Moses, T., Deng, W., & Zhang, L. (2009). *The use of two anchors in NEAT equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Pampel, F. C. (2000). Logistic Regression: A Primer. *Quantitative Applications in the Social Sciences Series* (no. 132). Thousand Oaks, CA: Sage.

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

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.

Peterson, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Academic Press.

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.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

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*(1), 53-71.

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

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology, 45,* 81-90.

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.

Slakter, M. J. (1968a). The penalty for not guessing. *Journal of Educational Measurement, 5,* 141-144.

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

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (ETS Research Report RR-88-41). Princeton, NJ: Educational Testing Service.

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

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

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*(1), 17-30.

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

von Davier, A.A., Holland, P.W., & Thayer, D.T. (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*(1), 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. (2005). *A didactic approach to the use of IRT true-score equating* (ETS Research Report RR-05-26). Princeton, NJ: Educational Testing Service.

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.

Votaw, D. F. (1936). The effect of do not guess directions upon the validity of true-false or multiple-choice tests. *Journal of Educational Psychology, 27*, 698-703.

Walker, M. E., Allspach, J. R., & Liu, J. (2004). *Scaling the new SAT Writing section: Finding the best solution* (ETS Statistical Report SR-2004-61). Princeton, NJ: Educational Testing Service.

Wang, T., Lee, W.-C., 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*(8), 632-651.

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

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

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

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

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72*(1), 69-91.

Ziller, R. C. (1957). A measure of the gambling response-set in objective tests. *Psychometrika, 22*(3), 289-292.